

# An Efficient Algorithm for the Equation Tree Automaton *via* the $k$ -C-Continuations

Ludovic Mignot, Nadia Ouali Sebti and Djelloul Ziadi \*

Laboratoire LITIS - EA 4108 Université de Rouen, Avenue de l'Université  
76801 Saint-Étienne-du-Rouvray Cedex.

**Abstract.** Champarnaud and Ziadi, and Khorsi *et al.* show how to compute the equation automaton of word regular expression  $E$  *via* the  $k$ -C-Continuations. Kuske and Meinecke extend the computation of the equation automaton to a regular tree expression  $E$  over a ranked alphabet  $\Sigma$  and produce a  $O(R \cdot |E|^2)$  time and space complexity algorithm, where  $R$  is the maximal rank of a symbol occurring in  $\Sigma$  and  $|E|$  is the size of  $E$ . In this paper, we give a full description of the algorithm based on the acyclic minimization of Revuz. Our algorithm, which is performed in an  $O(|Q| \cdot |E|)$  time and space complexity, where  $|Q|$  is the number of states of the produced automaton, is more efficient than the one obtained by Kuske and Meinecke.

## 1 Introduction

Regular expressions, which are finite representatives of potentially infinite languages, are widely used in various application areas such as XML Schema Languages [10], logic and verification [14], *etc.* The concept of word regular expressions has been extended to tree regular expressions. Similarly to word expressions, one can convert them into finite recognizers, the tree automata.

The study of the different ways of conversion of regular expressions into automata and *vice versa* is a very active field. There exists a lot of techniques to transform regular expressions (resp. regular tree expressions) into finite automata [2,6,7,15] (resp. into finite tree automata [8,9]). As far as tree automata are concerned, computation algorithms are extensions of word cases. In [9], the computation of the position tree automaton from a regular tree expression has been achieved by extending the classical notions of Glushkov functions defined in [6], leading to the computation of an automaton which number of states is linear w.r.t. the number of occurrences of symbols but which number of transitions can be exponential. In the same paper, it is proved that this automaton can be reduced into a quadratic size recognizer.

On the other side, Kuske and Meinecke have extended the notion of word partial derivatives [1] into tree partial derivatives. They also present how to compute them extending from words to trees [8] the  $k$ -C-Continuation algorithm by Champarnaud and Ziadi [3]. They obtain an algorithm with  $O(R \cdot |E| \cdot |E|)$

---

\* {ludovic.mignot, nadia.ouali-sebti, djelloul.ziadi}@univ-rouen.fr

space and time complexity where  $R$  is the maximal rank of a symbol occurring in the finite ranked alphabet  $\Sigma$  and  $|E|$  is the size of the regular expression.

In this paper, we show how to extend a notion of  $k$ -C-Continuation in order to compute from a regular tree expression its equation tree automaton with an  $O(|E| + |Q| \cdot |E|)$  time and space complexity where  $|Q|$  is the number of its states. This constitutes an improvement in comparison with Kuske and Meinecke algorithm [8]. The paper is organized as follows: Section 2 outlines finite tree automata over ranked trees, regular tree expressions, and linearized regular tree expressions which allows the set of positions to be defined. Next, in Section 3 the notions of derivation and partial derivative of regular expression and set of regular expressions are introduced. Thus the definitions of equation tree automaton and  $k$ -C-Continuation tree automaton associated with the regular expression  $E$  is obtained. Afterwards, in Section 4 we present our algorithm which builds the equation tree automaton with an  $O(|E| + |Q| \cdot |E|)$  time and space complexity. Finally, Section 5 provides a full example of our construction.

## 2 Preliminaries

Let  $(\Sigma, \text{ar})$  be a *ranked alphabet*, where  $\Sigma$  is a finite set and  $\text{ar}$  represents the *rank* of  $\Sigma$  which is a mapping from  $\Sigma$  into  $\mathbb{N}$ . The set of symbols of rank  $n$  is denoted by  $\Sigma_n$ . The elements of rank 0 are called *constants*. A *tree*  $t$  over  $\Sigma$  is inductively defined as follows:  $t = a$ ,  $t = f(t_1, \dots, t_k)$  where  $a$  is any symbol in  $\Sigma_0$ ,  $k$  is any integer satisfying  $k \geq 1$ ,  $f$  is any symbol in  $\Sigma_k$  and  $t_1, \dots, t_k$  are any  $k$  trees over  $\Sigma$ . We denote by  $T_\Sigma$  the set of trees over  $\Sigma$ . A *tree language* is a subset of  $T_\Sigma$ . Let  $\Sigma_{\geq 1} = \Sigma \setminus \Sigma_0$  denote the set of *non-constant symbols* of the ranked alphabet  $\Sigma$ . A *Finite Tree Automaton* (FTA) [5,8]  $\mathcal{A}$  is a tuple  $(Q, \Sigma, Q_T, \Delta)$  where  $Q$  is a finite set of states,  $Q_T \subset Q$  is the set of *final states* and  $\Delta \subset \bigcup_{n \geq 0} (Q \times \Sigma_n \times Q^n)$  is the set of *transition rules*. This set is equivalent to the function  $\Delta$  from  $Q^n \times \Sigma_n \rightarrow 2^Q$  defined by  $(q, f, q_1, \dots, q_n) \in \Delta \Leftrightarrow q \in \Delta(q_1, \dots, q_n, f)$ . The domain of this function can be extended to  $(2^Q)^n \times \Sigma_n \rightarrow 2^Q$  as follows:  $\Delta(Q_1, \dots, Q_n, f) = \bigcup_{(q_1, \dots, q_n) \in Q_1 \times \dots \times Q_n} \Delta(q_1, \dots, q_n, f)$ . Finally, we denote by  $\Delta^*$  the function from  $T_\Sigma \rightarrow 2^Q$  defined for any tree in  $T_\Sigma$  as follows:

$$\Delta^*(t) = \begin{cases} \Delta(a) & \text{if } t = a, a \in \Sigma_0 \\ \Delta(f, \Delta^*(t_1), \dots, \Delta^*(t_n)) & \text{if } t = f(t_1, \dots, t_n), f \in \Sigma_n, t_1, \dots, t_n \in T_\Sigma \end{cases}$$

A tree is *accepted* by  $\mathcal{A}$  if and only if  $\Delta^*(t) \cap Q_T \neq \emptyset$ . The *language recognized* by  $\mathcal{L}(\mathcal{A})$  is the set of trees accepted by  $\mathcal{A}$  i.e.  $\mathcal{L}(\mathcal{A}) = \{t \in T_\Sigma \mid \Delta^*(t) \cap Q_T \neq \emptyset\}$ . A state  $q \in Q$  is *coaccessible* if  $q \in Q_T$  or if  $\exists Q' = \{q_1, \dots, q_n\} \subset Q$ ,  $f \in \Sigma_n$ ,  $q'$  a coaccessible state in  $Q$  such that  $q \in Q'$  and  $q' \in \Delta(f, q_1, \dots, q_n)$ . The *coaccessible part* of the automaton  $\mathcal{A}$  is the tree automaton  $\mathcal{A}' = (Q', \Sigma, \Delta', Q_T')$  where  $Q' = \{q \in Q \mid q \text{ is coaccessible}\}$  and  $\Delta' = \{(q, f, q_1, \dots, q_n) \in \Delta \mid \{q, q_1, \dots, q_n\} \subset Q'\}$ . It is easy to show that  $\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}')$ .

Let  $\sim$  be an equivalence relation over  $Q$ . We denote by  $[q]$  the equivalence class of any state  $q$  in  $Q$ . The *quotient* of  $\mathcal{A}$  w.r.t.  $\sim$  is the tree automaton  $\mathcal{A}_{/\sim} = (Q_{/\sim}, \Sigma, Q_{T/\sim}, \Delta_{/\sim})$  where:  $Q_{/\sim} = \{[q] \mid q \in Q\}$ ,  $Q_{T/\sim} = \{[q] \mid q \in Q_T\}$ ,  $\Delta_{/\sim} = \{([q], f, [q_1], \dots, [q_n]) \mid (q, f, q_1, \dots, q_n) \in \Delta\}$ .

For any integer  $n \geq 0$ , for any  $n$  languages  $L_1, \dots, L_n \subset T_\Sigma$ , and for any symbol  $f \in \Sigma_n$ ,  $f(L_1, \dots, L_n)$  is the tree language  $\{f(t_1, \dots, t_n) \mid t_i \in L_i\}$ . The *tree substitution* of a constant  $c$  in  $\Sigma$  by a language  $L \subset T_\Sigma$  in a tree  $t \in T_\Sigma$ , denoted by  $t\{c \leftarrow L\}$ , is the language inductively defined by  $L$  if  $t = c$ ;  $\{d\}$  if  $t = d$  where  $d \in \Sigma_0 \setminus \{c\}$ ;  $f(t_1\{c \leftarrow L\}, \dots, t_n\{c \leftarrow L\})$  if  $t = f(t_1, \dots, t_n)$  with  $f \in \Sigma_n$  and  $t_1, \dots, t_n$  any  $n$  trees over  $\Sigma$ . Let  $c$  be a symbol in  $\Sigma_0$ . The *c-product*  $L_1 \cdot_c L_2$  of two languages  $L_1, L_2 \subset T_\Sigma$  is defined by  $L_1 \cdot_c L_2 = \bigcup_{t \in L_1} \{t\{c \leftarrow L_2\}\}$ . The *iterated c-product* is inductively defined for  $L \subset T_\Sigma$  by:  $L^{0_c} = \{c\}$  and  $L^{(n+1)_c} = L^{n_c} \cup L \cdot_c L^{n_c}$ . The *c-closure* of  $L$  is defined by  $L^{*c} = \bigcup_{n \geq 0} L^{n_c}$ .

A *regular expression* over a ranked alphabet  $\Sigma$  is inductively defined by  $E \in \Sigma_0$ ,  $E = f(E_1, \dots, E_n)$ ,  $E = (E_1 + E_2)$ ,  $E = (E_1 \cdot_c E_2)$ ,  $E = (E_1^{*c})$ , where  $c \in \Sigma_0$ ,  $n \in \mathbb{N}$ ,  $f \in \Sigma_n$  and  $E_1, E_2, \dots, E_n$  are any  $n$  regular expressions over  $\Sigma$ . Parenthesis can be omitted when there is no ambiguity. We write  $E_1 = E_2$  if  $E_1$  and  $E_2$  graphically coincide. We denote by  $\text{RegExp}(\Sigma)$  the set of all regular expressions over  $\Sigma$ . Every regular expression  $E$  can be seen as a tree over the ranked alphabet  $\Sigma \cup \{+, \cdot_c, *c\}$  with  $c \in \Sigma_0$  where  $+$  and  $\cdot_c$  can be seen as a symbol of rank 2 and  $*c$  has rank 1. This tree is the syntax-tree  $T_E$  of  $E$ . The *alphabetical width*  $\|E\|$  of  $E$  is the number of occurrences of symbols of  $\Sigma$  in  $E$ . The *size*  $|E|$  of  $E$  is the size of its syntax tree  $T_E$ . The *language*  $\llbracket E \rrbracket$  denoted by  $E$  is inductively defined as  $\llbracket c \rrbracket = \{c\}$ ,  $\llbracket f(E_1, E_2, \dots, E_n) \rrbracket = f(\llbracket E_1 \rrbracket, \dots, \llbracket E_n \rrbracket)$ ,  $\llbracket E_1 + E_2 \rrbracket = \llbracket E_1 \rrbracket \cup \llbracket E_2 \rrbracket$ ,  $\llbracket E_1 \cdot_c E_2 \rrbracket = \llbracket E_1 \rrbracket \cdot_c \llbracket E_2 \rrbracket$ ,  $\llbracket E_1^{*c} \rrbracket = \llbracket E_1 \rrbracket^{*c}$  where  $n \in \mathbb{N}$ ,  $E_1, E_2, \dots, E_n$  are any  $n$  regular expressions,  $f \in \Sigma_n$  and  $c \in \Sigma_0$ . It is well known that a tree language is accepted by some tree automaton if and only if it can be denoted by a regular expression [5,8]. A regular expression  $E$  defined over  $\Sigma$  is *linear* if and only if every symbol of  $\Sigma_{\geq 1}$  appears at most once in  $E$ . Note that any constant symbol may occur more than once. Let  $E$  be a regular expression over  $\Sigma$ . The *linearized regular expression*  $\bar{E}^E$  in  $E$  of a regular expression  $E$  is obtained from  $E$  by marking differently all symbols of a rank greater than or equal to 1 (symbols of  $\Sigma_{\geq 1}$ ). The set of *marked symbols* with symbols of  $\Sigma_0$  is the ranked alphabet containing symbols called *positions*. We denote this set by  $\text{Pos}_E(E)$ . When there is no ambiguity we denote by  $\bar{F}$  the subexpression  $\bar{F}^E$  with  $F$  is a subexpression of  $E$ . The mapping  $h$  is defined from  $\text{Pos}_E(E)$  to  $\Sigma$  with  $h(\text{Pos}_E(E)_m) \subset \Sigma_m$  for every  $m \in \mathbb{N}$ . It associates with a marked symbol  $f_j \in \text{Pos}_E(E)_{\geq 1}$  the symbol  $f \in \Sigma_{\geq 1}$  and for a symbol  $c \in \Sigma_0$  the symbol  $h(c) = c$ . We can extend the mapping  $h$  naturally to  $\text{RegExp}(\text{Pos}_E(E)) \rightarrow \text{RegExp}(\Sigma)$  by  $h(a) = a$ ,  $h(E_1 + E_2) = h(E_1) + h(E_2)$ ,  $h(E_1 \cdot_c E_2) = h(E_1) \cdot_c h(E_2)$ ,  $h(E_1^{*c}) = h(E_1)^{*c}$ ,  $h(f_j(E_1, \dots, E_n)) = f(h(E_1), \dots, h(E_n))$ , with  $n \in \mathbb{N}$ ,  $a \in \Sigma_0$ ,  $f \in \Sigma_n$ ,  $f_j \in \text{Pos}_E(E)_n$  such that  $h(f_j) = f$  and  $E_1, \dots, E_n$  any regular expressions over  $\text{Pos}_E(E)$ .

*Example 1.* Let  $\Sigma_0 = \{a, c\}$ ,  $\Sigma_1 = \{g, h\}$ ,  $\Sigma_2 = \{f\}$  and  $\Sigma = \Sigma_0 \cup \Sigma_1 \cup \Sigma_2$  be a ranked alphabet. Let  $E, F, G$  be the three following regular expressions over  $\Sigma$ :  $F = ((c + a) + (g(c))^{*c})^{*c}$ ,  $G = f(a, h(c))$  and  $E = F \cdot_c G$ . The linearized forms of  $E$  and  $G$  are:  $\bar{E}^E = ((c + a) + (g_1(c))^{*c})^{*c} \cdot_c f_2(a, h_3(c))$ ,  $\bar{G}^G = f_1(a, h_2(c))$ .

The linearized form of  $G$  in  $E$  is  $\overline{G}^E = f_2(a, h_3(c))$ . Notice that  $\text{Pos}_G(G) = \{a, f_1, h_2\} \neq \text{Pos}_E(G) = \{a, f_2, h_3\}$ .

### 3 Tree Automata Computations

In this section, we recall how to compute from a regular expression  $E$  a tree automaton that accepts  $\llbracket E \rrbracket$ . We first recall the computation of the equation automaton  $\mathcal{A}_E$  of  $E$ , then we define the  $k$ -c-continuation automaton  $\mathcal{C}_E$ .

#### 3.1 The Equation Tree Automaton

In [8], Kuske and Meinecke extend the notion of word partial derivatives [1] to tree partial derivatives in order to compute from a regular expression  $E$  a tree automaton recognizing  $\llbracket E \rrbracket$ . Due to the notion of ranked alphabet, partial derivatives are no longer sets of expressions, but sets of tuples of expressions.

Let  $\mathcal{N} = (E_1, \dots, E_n)$  be a tuple of regular expressions,  $F$  be some regular expression and  $c \in \Sigma_0$ . Then  $\mathcal{N} \cdot_c F$  is the tuple  $(E_1 \cdot_c F, \dots, E_n \cdot_c F)$ . For  $\mathcal{S}$  a set of tuples of regular expressions,  $\mathcal{S} \cdot_c F$  is the set  $\mathcal{S} \cdot_c F = \{\mathcal{N} \cdot_c F \mid \mathcal{N} \in \mathcal{S}\}$ . Finally,  $\text{SET}(\mathcal{N}) = \{E_1, \dots, E_n\}$  and  $\text{SET}(\mathcal{S}) = \bigcup_{\mathcal{N} \in \mathcal{S}} \text{SET}(\mathcal{N})$ .

**Definition 1 ([8]).** Let  $E$  be a regular expression over a ranked alphabet  $\Sigma$  and  $f$  be a symbol in  $\Sigma_m$  with  $m \geq 1$  an integer. The set  $f^{-1}(E)$  of tuples of regular expressions is defined as follows:

$$\begin{aligned} f^{-1}(g(E_1, \dots, E_n)) &= \begin{cases} \{(E_1, \dots, E_n)\} & \text{if } f = g \\ \emptyset & \text{otherwise} \end{cases} \\ f^{-1}(F + G) &= f^{-1}(F) \cup f^{-1}(G) \\ f^{-1}(F \cdot_c G) &= \begin{cases} f^{-1}(F) \cdot_c G & \text{if } c \notin \llbracket F \rrbracket \\ f^{-1}(F) \cdot_c G \cup f^{-1}(G) & \text{otherwise} \end{cases} \\ f^{-1}(F^{*c}) &= f^{-1}(F) \cdot_c F^{*c} \end{aligned}$$

The function  $f^{-1}$  is extended to any set  $S$  of regular expressions as follows:  
 $f^{-1}(S) = \bigcup_{E \in S} f^{-1}(E)$ .

The *partial derivative* of  $E$  w.r.t. a word  $w \in \Sigma_{\geq 1}^*$ , denoted by  $\partial_w(E)$ , is the set of regular expressions inductively defined by:

$$\partial_w(E) = \begin{cases} \{E\} & \text{if } w = \varepsilon \\ \text{SET}(f^{-1}(\partial_u(E))) & \text{if } w = uf, f \in \Sigma_{\geq 1}, u \in \Sigma_{\geq 1}^* \end{cases}$$

The partial derivation is extended to any subset  $U$  of  $\Sigma_{\geq 1}^*$  as by  $\partial_U(E) = \bigcup_{w \in U} \partial_w(E)$ . Note that  $\partial_{uf}(E) = \partial_f(\partial_u(E)) = \bigcup_{F \in \partial_u(E)} \partial_f(F)$ .

**Definition 2.** Let  $E$  be a regular expression over a ranked alphabet  $\Sigma$ . The Equation Automaton of  $E$  is the tree automaton  $\mathcal{A}_E = (Q, \Sigma, Q_T, \Delta)$  defined by  $Q = \partial_{\Sigma_{\geq 1}^*}(E)$ ,  $Q_T = \{E\}$ , and

$$\Delta = \left\{ \begin{aligned} &\{(F, f, G_1, \dots, G_m) \mid F \in Q, f \in \Sigma_m, m \geq 1, (G_1, \dots, G_m) \in f^{-1}(F)\} \\ &\cup \{(F, c) \mid c \in (\llbracket F \rrbracket \cap \Sigma_0)\} \end{aligned} \right\}$$

**Theorem 1 ([8]).** Let  $E$  be a regular expression and  $\mathcal{A}_E$  be the equation tree automaton associated with  $E$ . Then  $\mathcal{L}(\mathcal{A}_E) = \llbracket E \rrbracket$ .

### 3.2 The C-Continuation Tree Automaton

In [8], Kuske and Meinecke show how to efficiently compute the equation tree automaton of a regular expression *via* an extension of Champarnaud and Ziadi's  $k$ -C-Continuation [3,4,7]. In this section, we show how to inductively compute them. The main difference with [8] is that the  $k$ -c-continuations are here computed using alternative formulae, and not using the partial derivation. As a consequence, any symbol that appears in the expression  $E$  admits a non-empty  $k$ -c-continuation (*e.g.* in [8], there is no continuation for  $g$  in  $E = a \cdot_b g(c)$ ).

**Definition 3.** Let  $E$  be linear. Let  $k$  and  $m$  be two integers such that  $1 \leq k \leq m$ . Let  $f$  be in  $(\Sigma_E \cap \Sigma_m)$ . The  $k$ -C-continuation  $C_{f^k}(E)$  of  $f$  in  $E$  is the regular expression defined by:

$$\begin{aligned} C_{f^k}(g(E_1, \dots, E_m)) &= \begin{cases} E_k & \text{if } f = g \\ C_{f^k}(E_j) & \text{if } f \in \Sigma_{E_j} \end{cases} \\ C_{f^k}(F + G) &= \begin{cases} C_{f^k}(F) & \text{if } f \in \Sigma_F \\ C_{f^k}(G) & \text{if } f \in \Sigma_G \end{cases} \\ C_{f^k}(F \cdot_c G) &= \begin{cases} C_{f^k}(F) \cdot_c G & \text{if } f \in \Sigma_F \\ C_{f^k}(G) & \text{otherwise} \end{cases} \\ C_{f^k}(F^{*c}) &= C_{f^k}(F) \cdot_c F^{*c} \end{aligned}$$

By convention, we set  $C_{\varepsilon^1}(E) = E$ .

Let us first show the relation between partial derivation and  $k$ -c-continuation.

**Lemma 1.** Let  $E$  be linear,  $n$ ,  $m$  and  $k$  be three integers such that  $n, m \geq 1$ ,  $1 \leq k \leq m$ ,  $f \in \Sigma_n$  and  $g \in \Sigma_m \cup \{\varepsilon\}$ . If  $f^{-1}(C_{g^k}(E)) \neq \emptyset$  then  $f^{-1}(C_{g^k}(E)) = \{(C_{f^1}(E), \dots, C_{f^n}(E))\}$ .

*Proof.* By induction over the structure of  $E$ . For any symbol  $g \in \Sigma_p \cup \{\varepsilon\}$  and for any expression  $F$ , let us set  $C_g(F) = (C_{g^1}(F), \dots, C_{g^p}(F))$ .

1. Let us suppose that  $E = h(E_1, \dots, E_m)$ . Three cases have to be considered:
  - (a) If  $g = \varepsilon$ , then  $k = 1$  and  $f^{-1}(C_{g^k}(E)) = f^{-1}(E)$ . Since  $f^{-1}(C_{g^k}(E)) \neq \emptyset$ ,  $f = h$ . Hence,  $f^{-1}(E) = \{(E_1, \dots, E_m)\}$ . Moreover, for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = E_j$ . Consequently,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
  - (b) Let us suppose that  $g \neq \varepsilon$  and  $g \neq h$ . Hence  $C_{g^k}(E) = C_{g^k}(E_l)$  with  $f \in \Sigma_{E_l}$ . By induction hypothesis,  $f^{-1}(C_{g^k}(E_j)) = \{C_f(E_l)\}$ . Moreover, for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_l)$ . Consequently,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
  - (c) Let us suppose that  $g \neq \varepsilon$  and  $g = h$ . Hence  $C_{g^k}(E) = E_k$ . Since  $f^{-1}(C_{g^k}(E)) \neq \emptyset$ , then  $f \in \Sigma_{E_k}$ . Thus,  $f \neq h$ . By definition,  $E_k = C_{\varepsilon^1}(E_k)$ . By induction hypothesis,  $f^{-1}(C_{\varepsilon^1}(E_k)) = \{C_f(E_k)\}$ . Since  $f \neq h$  and since  $f \in \Sigma_{E_k}$ , for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_k)$ . Consequently,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
2. Suppose that  $E = E_1 + E_2$ . Suppose that  $f \in \Sigma_{E_1}$ . Then  $f^{-1}(C_{g^k}(E)) = f^{-1}(C_{g^k}(E_1))$ . By induction hypothesis,  $f^{-1}(C_{g^k}(E_1)) = \{C_f(E_1)\}$ . Finally, since for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_1)$ , it holds  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ . The prove is identical whenever  $f \in \Sigma_{E_2}$ .

3. Let us suppose that  $E = E_1 \cdot_c E_2$ . Two cases have to be considered:
  - (a) If  $g \in \Sigma_{E_1}$ ,  $f^{-1}(C_{g^k}(E)) = f^{-1}(C_{g^k}(E_1) \cdot_c E_2)$ . If  $f \in \Sigma_{E_1}$ , then  $f^{-1}(C_{g^k}(E_1) \cdot_c E_2) = f^{-1}(C_{g^k}(E_1)) \cdot_c E_2$ ; otherwise,  $f^{-1}(C_{g^k}(E_1) \cdot_c E_2) = f^{-1}(E_2)$ . Hence, according to induction hypothesis, either  $f^{-1}(C_{g^k}(E_1) \cdot_c E_2) = \{(C_{f^1}(E_1) \cdot_c E_2, \dots, C_{f^n}(E_1) \cdot_c E_2)\}$ , or  $f^{-1}(C_{g^k}(E_1) \cdot_c E_2) = \{(C_{f^1}(E_2), \dots, C_{f^n}(E_2))\}$ . By definition, considering whether  $f \in \Sigma_{E_1}$ , for any integer  $1 \leq j \leq n$ , either  $C_{f^j}(E) = C_{f^j}(E_1) \cdot_c E_2$  or  $C_{f^j}(E) = C_{f^j}(E_2)$ . In both of these cases,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
  - (b) If  $g \in \Sigma_{E_2}$ ,  $f^{-1}(C_{g^k}(E)) = f^{-1}(C_{g^k}(E_2))$ . By induction hypothesis,  $f^{-1}(C_{g^k}(E_2)) = \{(C_{f^1}(E_2), \dots, C_{f^n}(E_2))\}$ . Moreover, for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_2)$ . Consequently,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
4. Let us suppose that  $E = E_1^{*c}$ . Two cases have to be considered:
  - (a) If  $g = \varepsilon$ , then  $f^{-1}(C_{g^k}(E)) = f^{-1}(E_1^{*c}) = f^{-1}(E_1) \cdot_c E_1^{*c}$ . By definition,  $E_1 = C_\varepsilon(E_1)$ . Hence by induction hypothesis,  $f^{-1}(C_\varepsilon(E_1)) \cdot_c E_1^{*c} = \{(C_{f^1}(E_1) \cdot_c E_1^{*c}, \dots, C_{f^n}(E_1) \cdot_c E_1^{*c})\}$ . Moreover, for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_1) \cdot_c E_1^{*c}$ . Consequently,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .
  - (b) Suppose that  $g \neq \varepsilon$ . Then  $f^{-1}(C_{g^k}(E)) = f^{-1}(C_{g^k}(E_1) \cdot_c E_1^{*c})$ . Depending whether  $c$  belongs to  $\llbracket C_{g^k}(E_1) \rrbracket$ , either  $f^{-1}(C_{g^k}(E_1) \cdot_c E_1^{*c}) = f^{-1}(C_{g^k}(E_1)) \cdot_c E_1^{*c}$  or  $f^{-1}(C_{g^k}(E_1) \cdot_c E_1^{*c}) = f^{-1}(C_{g^k}(E_1)) \cdot_c E_1^{*c} \cup f^{-1}(E_1^{*c})$ . Since  $E_1^{*c} = C_\varepsilon^1(E_1^{*c})$ , it holds by induction hypothesis that either  $f^{-1}(C_{g^k}(E_1) \cdot_c E_1^{*c}) = \{(C_{f^1}(E_1) \cdot_c E_1^{*c}, \dots, C_{f^n}(E_1) \cdot_c E_1^{*c})\}$  or  $f^{-1}(C_{g^k}(E_1) \cdot_c E_1^{*c}) = \{(C_{f^1}(E_1) \cdot_c E_1^{*c}, \dots, C_{f^n}(E_1) \cdot_c E_1^{*c}) \cup \{C_f(E_1^{*c})\}\}$ . Finally, since for any integer  $1 \leq j \leq n$ ,  $C_{f^j}(E) = C_{f^j}(E_1) \cdot_c E_1^{*c}$ , in both of these cases,  $f^{-1}(C_{g^k}(E)) = \{C_f(E)\}$ .

□

**Proposition 1.** *Let  $E$  be linear and  $f \in \Sigma_n$  with  $n \geq 1$ . Let  $u$  be a word in  $\Sigma_{\geq 1}^*$ . If  $f^{-1}(\partial_u(E)) \neq \emptyset$  then  $f^{-1}(\partial_u(E)) = \{(C_{f^1}(E), \dots, C_{f^n}(E))\}$ .*

*Proof.* By recurrence over the length of  $u$ . For any symbol  $g \in \Sigma_p \cup \{\varepsilon\}$  and for any expression  $F$ , let us set  $C_g(F) = (C_{g^1}(F), \dots, C_{g^p}(F))$ .

1. Let  $u = \varepsilon$ . Then  $f^{-1}(\partial_u(E)) = f^{-1}(E)$ . By definition,  $f^{-1}(E) = f^{-1}(C_\varepsilon^1(E))$ . According to Lemma 1,  $f^{-1}(C_\varepsilon^1(E)) = \{C_f(E)\}$ .
2. Let  $u = wg$  with  $w$  a word in  $\Sigma_{\geq 1}^*$  and  $g$  a symbol in  $\Sigma_m$ . Then  $f^{-1}(\partial_u(E)) = f^{-1}(SET(g^{-1}(\partial_u(E))))$ . According to recurrence hypothesis, it holds that  $SET(g^{-1}(\partial_u(E))) = SET(\{C_g(E)\}) = \{C_{g^1}(E), \dots, C_{g^m}(E)\}$ . By definition,  $f^{-1}(\{C_{g^1}(E), \dots, C_{g^m}(E)\}) = \bigcup_{1 \leq i \leq m} f^{-1}(C_{g^i}(E))$ . According to Lemma 1, for any integer  $i$  such that  $f^{-1}(C_{g^i}(E)) \neq \emptyset$ , it holds  $f^{-1}(C_{g^i}(E)) = \{(C_{f^1}(E), \dots, C_{f^n}(E))\}$ . Since  $f^{-1}(\partial_u(E)) \neq \emptyset$ , there exists at least one integer  $i$  such that  $f^{-1}(C_{g^i}(E)) \neq \emptyset$ . Consequently,  $\bigcup_{1 \leq i \leq m} f^{-1}(C_{g^i}(E)) = \{C_f(E)\}$ .

□

**Definition 4.** *The automaton  $\bar{\mathcal{C}}_E = (Q_{\bar{\mathcal{C}}}, \text{Pos}_E(E), \{C_{\varepsilon^1}(\bar{E})\}, \Delta_{\bar{\mathcal{C}}})$  is defined by*

$$- Q_{\bar{\mathcal{C}}} = \{C_{f_j^k}(\bar{E}) \mid f_j \in \text{Pos}_E(E)_m, 1 \leq k \leq m\} \cup \{C_{\varepsilon^1}(\bar{E})\},$$

$$- \Delta_{\overline{\mathcal{C}}} = \left\{ \begin{array}{l} \{(C_x(\overline{\mathbf{E}}), g_i, \mathfrak{C}_{g_i}) \mid g_i \in \text{Pos}_{\mathbf{E}}(\mathbf{E})_m, m \geq 1, \mathfrak{C}_{g_i} \in g_i^{-1}(C_x(\overline{\mathbf{E}}))\} \\ \cup \{(C_x(\overline{\mathbf{E}}), c) \mid c \in \llbracket C_x(\overline{\mathbf{E}}) \rrbracket \cap \Sigma_0\} \end{array} \right\}$$

where for any symbol  $g_i$  in  $\text{Pos}_{\mathbf{E}}(\mathbf{E})_m$ ,  $\mathfrak{C}_{g_i} = (C_{g_i^1}(\overline{\mathbf{E}}), \dots, C_{g_i^m}(\overline{\mathbf{E}}))$ .

The following lemma illustrates the link between  $\overline{\mathcal{C}}_{\mathbf{E}}$  and  $\mathcal{A}_{\overline{\mathbf{E}}}$ .

**Lemma 2.** *The coaccessible part of  $\overline{\mathcal{C}}_{\mathbf{E}}$  is equal to  $\mathcal{A}_{\overline{\mathbf{E}}}$ .*

*Proof.* The expression  $\overline{\mathbf{E}}$  is the final state of the two automata. Let us suppose now that  $q$  is a coaccessible state both in  $\overline{\mathcal{C}}_{\mathbf{E}}$  and  $\mathcal{A}_{\overline{\mathbf{E}}}$ . Hence, from Definition 4 and from Definition 2:

$$\begin{aligned} & \text{there exists a transition } (q, f, q_1, \dots, q_n) \text{ in } \mathcal{A}_{\overline{\mathbf{E}}} \\ \Leftrightarrow & (q_1, \dots, q_n) \in f^{-1}(q) \\ \Leftrightarrow & (q_1, \dots, q_n) = (C_f^1(\overline{\mathbf{E}}), \dots, C_f^n(\overline{\mathbf{E}})) \in f^{-1}(q) \text{ (Proposition 1)} \\ \Leftrightarrow & \text{there exists a transition } (q, f, q_1, \dots, q_n) \text{ in } \overline{\mathcal{C}}_{\mathbf{E}}. \end{aligned}$$

Hence, the states  $q_1, \dots, q_n$  are coaccessible from  $q$  by  $f$  in  $\overline{\mathcal{C}}_{\mathbf{E}}$  if and only if they are in  $\mathcal{A}_{\overline{\mathbf{E}}}$ . Consequently, the coaccessible part of  $\overline{\mathcal{C}}_{\mathbf{E}}$  is equal to the equation tree automaton  $\mathcal{A}_{\overline{\mathbf{E}}}$ .  $\square$

**Corollary 1.** *The automaton  $\overline{\mathcal{C}}_{\mathbf{E}}$  accepts  $\llbracket \overline{\mathbf{E}} \rrbracket$ .*

The *C-Continuation tree automaton*  $\mathcal{C}_{\mathbf{E}}$  associated with  $\mathbf{E}$  is obtained by replacing each transition  $(C_x(\overline{\mathbf{E}}), g_i, C_{g_i^1}(\overline{\mathbf{E}}), \dots, C_{g_i^m}(\overline{\mathbf{E}}))$  of the tree automaton  $\overline{\mathcal{C}}_{\mathbf{E}}$  by  $(C_x(\overline{\mathbf{E}}), h(g_i), C_{g_i^1}(\overline{\mathbf{E}}), \dots, C_{g_i^m}(\overline{\mathbf{E}}))$ .

**Corollary 2.**  $h(\mathcal{L}(\overline{\mathcal{C}}_{\mathbf{E}})) = \mathcal{L}(\mathcal{C}_{\mathbf{E}}) = \llbracket \mathbf{E} \rrbracket$ .

In what follows, for any two trees  $s$  and  $t$ , we denote by  $s \preccurlyeq t$  the relation "s is a subtree of t". Let  $k$  be an integer. We denote by  $\text{root}(s)$  the root of any tree  $s$  and by  $k\text{-child}(t)$ , for a tree  $t = f(t_1, \dots, t_n)$ , the  $k^{\text{th}}$  child of  $f$  in  $t$  that is root of  $t_k$  if it exists.

Let  $1 \leq k \leq m$  be two integers and  $f_j$  be a symbol in  $\text{Pos}_{\mathbf{E}}(\mathbf{E})_m$ . The sets  $\text{First}(\mathbf{E})$  is the subset of  $\text{Pos}_{\mathbf{E}}(\mathbf{E})$  defined by  $\text{First}(\mathbf{E}) = \{\text{root}(t) \in \text{Pos}_{\mathbf{E}}(\mathbf{E}) \mid t \in \llbracket \overline{\mathbf{E}} \rrbracket\}$ . The set  $\text{Follow}(\mathbf{E}, f_j, k)$  is the subset of  $\text{Pos}_{\mathbf{E}}(\mathbf{E})$  defined by  $\text{Follow}(\mathbf{E}, f_j, k) = \{g_i \in \text{Pos}_{\mathbf{E}}(\mathbf{E}) \mid \exists t \in \llbracket \overline{\mathbf{E}} \rrbracket, \exists s \preccurlyeq t, \text{root}(s) = f_j, k\text{-child}(s) = g_i\}$ .

**Proposition 2 ([9]).** *The computation of all the sets  $(\text{Follow}(\mathbf{E}, f_j, k))_{1 \leq k \leq m, f_j \in \text{Pos}_{\mathbf{E}}(\mathbf{E})_m}$  can be done with an  $O(\|\mathbf{E}\|)$  time and space complexity.*

**Proposition 3.** *Let  $1 \leq k \leq m$  be two integers and  $f_j$  be a position in  $\text{Pos}_{\mathbf{E}}(\mathbf{E})_m$ . If  $\text{Follow}(\mathbf{E}, f_j, k) \neq \emptyset$  then  $\text{Follow}(\mathbf{E}, f_j, k) = \text{First}(C_{f_j^k}(\overline{\mathbf{E}}))$ .*

*Proof.* Let  $\mathbf{E}$  be a linear regular expression over a ranked alphabet  $\Sigma$ ,  $1 \leq k \leq m$  be two integers and  $f$  be a symbol in  $\Sigma_m$ . The set  $\lambda^f(\mathbf{E}, k)$  is the subset of  $\Sigma_0$  defined by  $\lambda^f(\mathbf{E}, k) = \{c \in \Sigma_0 \mid \exists t \in \llbracket \mathbf{E} \rrbracket, \exists f(t_1, \dots, t_m) \preccurlyeq t, t_k = c\}$ . The set  $\lambda(\mathbf{E})$  is the subset of  $\Sigma_0$  defined by  $\lambda(\mathbf{E}) = \bigcup_{g \in \Sigma_m, 1 \leq k \leq m} \lambda^g(\mathbf{E}, k)$ .

Let  $\mathbf{E}$  be a regular expression over a ranked alphabet  $\Sigma$ ,  $1 \leq k \leq m$  be two integers and  $f_j$  be a symbol in  $\text{Pos}_{\mathbf{E}}(\mathbf{E})_m$ . In [9], it is shown, using alternative

and equivalent formulae, that the set  $\text{Follow}(E, f_j, k)$  is equal to  $\text{Follow}(\overline{E}, f_j, k)$ , where  $\text{Follow}(F, f_j, k)$  is the subset of  $\text{Pos}_E(E)$  inductively defined for any linear regular expression  $F$  as follows:

$$\begin{aligned}
& \text{Follow}(a, f, k) = \emptyset, \\
& \text{Follow}(E_1 + E_2, f, k) = \begin{cases} \text{Follow}(E_1, f, k) & \text{if } f \in \Sigma_{E_1}, \\ \text{Follow}(E_2, f, k) & \text{if } f \in \Sigma_{E_2}, \end{cases} \\
& \text{Follow}(E_1 \cdot_c E_2, f, k) = \begin{cases} (\text{Follow}(E_1, f, k) \setminus \{c\}) \cup \text{First}(E_2) & \text{if } c \in \lambda^f(E_1, k), \\ \text{Follow}(E_1, f, k) & \text{if } f \in \Sigma_{E_1} \wedge c \notin \lambda^f(E_1, k), \\ \text{Follow}(E_2, f, k) & \text{if } f \in \Sigma_{E_2} \wedge c \in \lambda(E_1), \\ \emptyset & \text{otherwise,} \end{cases} \\
& \text{Follow}(E_1^*, f, k) = \begin{cases} \text{Follow}(E_1, f, k) \cup \text{First}(E_1) & \text{if } c \in \lambda(E_1), \\ \text{Follow}(E_1, f, k) & \text{otherwise,} \end{cases} \\
& \text{Follow}(g(E_1, \dots, E_n), f, k) = \begin{cases} \text{First}(E_k) & \text{if } f = g, \\ \text{Follow}(E_l, f, k) & \text{if } f \in \Sigma_{E_l}. \end{cases}
\end{aligned}$$

Since by definition  $\text{Follow}(E, f_j, k) = \text{Follow}(\overline{E}, f_j, k)$ , let us show by induction over  $\overline{E}$  that if  $\text{Follow}(E, f_j, k) \neq \emptyset$  then  $\text{Follow}(E, f_j, k) = \text{First}(C_{f_j^k}(\overline{E}))$ .

Let us set  $\overline{E} = F$ .

Suppose that  $F = f_j(F_1, \dots, F_m)$ . Hence  $\text{Follow}(F, f_j, k) = \text{First}(F_k)$ . Moreover by definition  $C_{f_j^k}(F) = F_k$ . Then  $\text{Follow}(F, f_j, k) = \text{First}(C_{f_j^k}(F))$ . The property is true for the base case.

Assuming that the property holds for the subexpressions of  $F$ .

1. Consider that  $F = g_i(F_1, \dots, F_m)$  with  $f_j \neq g_i$ . Then by definition  $\text{Follow}(F, f_j, k) = \text{Follow}(F_l, f_j, k)$  with  $f_j \in \Sigma_{F_l}$ . By induction hypothesis,  $\text{Follow}(F_l, f_j, k) = \text{First}(C_{f_j^k}(F_l))$ . Moreover, from Definition 3,  $C_{f_j^k}(F) = C_{f_j^k}(F_l)$ . Consequently,  $\text{Follow}(F, f_j, k) = \text{First}(C_{f_j^k}(F))$ .
2. Let us consider that  $F = F_1 + F_2$ . Suppose that  $f_j \in \Sigma_{F_i}$  with  $i \in \{1, 2\}$ . Hence  $\text{Follow}(F_1 + F_2, f_j, k) = \text{Follow}(F_i, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_i, f_j, k) = \text{First}(C_{f_j^k}(F_i))$ . From Definition 3,  $C_{f_j^k}(F_1 + F_2) = C_{f_j^k}(F_i)$ . Consequently,  $\text{Follow}(F_1 + F_2, f_j, k) = \text{First}(C_{f_j^k}(F_1 + F_2))$ .
3. Consider that  $F = F_1 \cdot_c F_2$ . Three cases may occur.
  - (a) Suppose that  $c \in \lambda^{f_j}(F_1, k)$ . Then  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = (\text{Follow}(F_1, f_j, k) \setminus \{c\}) \cup \text{First}(F_2)$ . By induction hypothesis,  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = (\text{First}(C_{f_j^k}(F_1)) \setminus \{c\}) \cup \text{First}(F_2)$ . Moreover, from Definition 3,  $C_{f_j^k}(F) = C_{f_j^k}(F_1) \cdot_c F_2$ . Since  $c \in \lambda^{f_j}(F_1, k)$ , then by definition of  $\lambda^{f_j}(F_1, k)$ ,  $c \in \text{Follow}(F_1, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$  and then  $c \in \llbracket C_{f_j^k}(F_1) \rrbracket$ . Consequently, by definition,  $\text{First}(C_{f_j^k}(F_1) \cdot_c F_2) = (\text{First}(C_{f_j^k}(F_1)) \setminus \{c\}) \cup \text{First}(F_2)$ . Therefore,  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = \text{First}(C_{f_j^k}(F_1 \cdot_c F_2))$ .
  - (b) Consider that  $c \notin \lambda^{f_j}(F_1, k)$  and  $f_j \in \Sigma_{F_1}$ . In this case  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = \text{Follow}(F_1, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$ . From Definition 3,  $C_{f_j^k}(F_1 \cdot_c F_2) = C_{f_j^k}(F_1) \cdot_c F_2$ . Since  $c \notin \lambda^{f_j}(F_1, k)$ , then by definition  $c \notin \text{Follow}(F_1, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$  and then  $c \notin \llbracket C_{f_j^k}(F_1) \rrbracket$ .



- Consequently,  $\text{First}(C_{f_j^k}(F_1 \cdot_c F_2)) = \text{First}(C_{f_j^k}(F_1))$ . Then  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = \text{First}(C_{f_j^k}(F_1 \cdot_c F_2))$ .
- (c) Consider that  $c \in \lambda(F_1)$  and  $f_j \in \Sigma_{F_2}$ . In this case  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = \text{Follow}(F_2, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_2, f_j, k) = \text{First}(C_{f_j^k}(F_2))$ . From Definition 3,  $C_{f_j^k}(F_1 \cdot_c F_2) = C_{f_j^k}(F_2)$ . Therefore,  $\text{Follow}(F_1 \cdot_c F_2, f_j, k) = \text{First}(C_{f_j^k}(F_1 \cdot_c F_2))$ .
4. Consider that  $F = F_1^{*c}$ . By Definition 3,  $C_{f_j^k}(F_1^{*c}) = C_{f_j^k}(F_1) \cdot_c F_1$ . Two cases may occur.
- (a) Suppose that  $c \in \lambda(F_1)$ . In this case,  $\text{Follow}(F_1^{*c}, f_j, k) = \text{Follow}(F_1, f_j, k) \cup \text{First}(F_1^{*c})$ . By induction hypothesis,  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$ . By definition,  $c \in \text{Follow}(F_1, f_j, k)$ . Since by induction  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$ ,  $c \in \text{First}(C_{f_j^k}(F_1))$  and then  $c \in \llbracket C_{f_j^k}(F_1) \rrbracket$ . Consequently,  $\text{First}(C_{f_j^k}(F_1) \cdot_c F_1) = \text{First}(C_{f_j^k}(F_1)) \cup \text{First}(F_1)$ . Consequently  $\text{Follow}(F_1^{*c}, f_j, k) = \text{Follow}(F_1, f_j, k) \cup \text{First}(F_1^{*c}) = \text{First}(C_{f_j^k}(F_1)) \cup \text{First}(F_1) = \text{First}(C_{f_j^k}(F_1^{*c}))$ .
- (b) Suppose that  $c \notin \lambda(F_1)$ . In this case,  $\text{Follow}(F_1^{*c}, f_j, k) = \text{Follow}(F_1, f_j, k)$ . By induction hypothesis,  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$ . By definition,  $c \notin \text{Follow}(F_1, f_j, k)$ . Since by induction  $\text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1))$ ,  $c \notin \text{First}(C_{f_j^k}(F_1))$  and then  $c \notin \llbracket C_{f_j^k}(F_1) \rrbracket$ . Consequently,  $\text{First}(C_{f_j^k}(F_1) \cdot_c F_1) = \text{First}(C_{f_j^k}(F_1))$ . Consequently  $\text{Follow}(F_1^{*c}, f_j, k) = \text{Follow}(F_1, f_j, k) = \text{First}(C_{f_j^k}(F_1)) = \text{First}(C_{f_j^k}(F_1^{*c}))$ .

□

**Proposition 4.** Let  $1 \leq k \leq m$  be two integers,  $f_j$  be a symbol in  $\text{Pos}_E(E)_m$  and  $g_i$  be a symbol in  $\text{Pos}_E(E)$ . Then  $g_i^{-1}(C_{f_j^k}(\bar{E})) \neq \emptyset \Leftrightarrow g_i \in \text{First}(C_{f_j^k}(\bar{E}))$ .

*Proof.* Let  $F$  be a linear expression. Let us show by induction over the structure of  $F$  that  $g_i^{-1}(F) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F)$ .

1. Consider that  $F = g_i(F_1, \dots, F_n)$ . By definition,  $\text{First}(F) = \{g_i\}$ . By definition,  $g_i^{-1}(F) = \{(F_1, \dots, F_n)\}$ . Hence the two conditions are both satisfied.
2. Consider that  $F = f(F_1, \dots, F_n)$  with  $f \in \Sigma_F \setminus \{g_i\}$ . By definition,  $\text{First}(F) = \{f\}$ . By definition,  $g_i^{-1}(F) = \emptyset$ . Hence the two conditions are both unsatisfied.
3. If  $F = F_1 + F_2$ , then according to [9],  $\text{First}(F) = \text{First}(F_1) \cup \text{First}(F_2)$ . By definition,  $g_i^{-1}(F) = g_i^{-1}(F_1) \cup g_i^{-1}(F_2)$ . By induction hypothesis, for  $l \in \{1, 2\}$ ,  $g_i^{-1}(F_l) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F_l)$ . Consequently,  $g_i^{-1}(F) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F_1) \vee g_i \in \text{First}(F_2) \Leftrightarrow g_i \in \text{First}(F_1) \cup \text{First}(F_2)$ .
4. If  $F = F_1 \cdot_c F_2$ , then according to [9],  $\text{First}(F) = \text{First}(F_1) \cup (\text{First}(F_2) \mid c \in c \in \llbracket F_1 \rrbracket)$ . By definition,  $g_i^{-1}(F) = g_i^{-1}(F_1) \cdot_c F_2 \cup (g_i^{-1}(F_2) \mid c \in \llbracket F_1 \rrbracket)$ . By induction hypothesis, for  $l \in \{1, 2\}$ ,  $g_i^{-1}(F_l) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F_l)$ . Consequently,  $g_i^{-1}(F) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F)$ .
5. If  $F = F_1^{*c}$ , then according to [9],  $\text{First}(F) = \text{First}(F_1)$ . By definition,  $g_i^{-1}(F) = g_i^{-1}(F_1) \cdot_c F_1^{*c}$ . By induction hypothesis,  $g_i^{-1}(F_1) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F_1)$ . Consequently,  $g_i^{-1}(F) \neq \emptyset \Leftrightarrow g_i \in \text{First}(F)$ .

As a direct consequence, the conditions of Proposition 4 are equivalent.  $\square$

**Lemma 3.** *Let  $1 \leq k \leq m$  be two integers and  $f_j$  be a position in  $\text{Pos}_E(\bar{E})_m$ . If  $\text{Follow}(E, f_j, k) = \emptyset$  then  $C_{f_j^k}(\bar{E})$  is not a coaccessible state in  $\mathcal{C}_E$ .*

*Proof.* Let us first show that for any state  $q = C_{f_j^k}(\bar{E})$ , there exists a tree  $t$  such that  $q \in \Delta^*(t)$ , where  $\Delta$  is the transition function of  $\bar{\mathcal{C}}_E$  (proposition **P** in the following). By definition,  $\llbracket q \rrbracket$  is not empty. If there exists a constant  $c \in \llbracket q \rrbracket$ , then by construction  $q \in \Delta^*(c)$ . If  $t = g_i(t_1, \dots, t_n) \in \llbracket q \rrbracket$ , then by definition  $g_i \in \text{First}(q)$ . According to Proposition 4,  $g_i^{-1}(C_{f_j^k}(\bar{E})) \neq \emptyset$ . Furthermore, according to Lemma 1,  $g_i^{-1}(C_{f_j^k}(\bar{E})) = \{(C_{g_i^1}(\bar{E}), \dots, C_{g_i^n}(\bar{E}))\}$ . Hence, the states  $q_1 = C_{g_i^1}(\bar{E}), \dots, q_n = C_{g_i^n}(\bar{E})$  are coaccessible from  $q$ . By induction hypothesis, there exists a tree  $t'_l$  in  $\llbracket q_l \rrbracket$  such that  $q_l \in \Delta^*(t'_l)$ . As a direct consequence,  $q \in \Delta^*(g_i(t'_1, \dots, t'_n))$ .

Let us show that if  $q$  is coaccessible, then there exists a tree  $t$  in  $\mathcal{L}(\bar{\mathcal{C}}_E)$  for any tree  $s$  satisfying  $q \in \Delta^*(s)$  such that  $s \preceq t$  (proposition **P'** in the following). If  $q = C_\varepsilon^1(\bar{E})$ , any tree  $s$  such that  $s \in \Delta^*(t)$  is accepted since  $q$  is final. Setting  $t = s$ , property holds. Otherwise,  $q$  is coaccessible from a state  $p$ . By construction, there exists a transition  $(p, f_j, (q_1, \dots, q_m))$  with  $q_k = q$ . By induction hypothesis, there exists a tree  $t$  in  $\mathcal{L}(\bar{\mathcal{C}}_E)$  for any tree  $s'$  satisfying  $p \in \Delta^*(s')$  such that  $s' \preceq t$ . Since any tree  $s$  satisfying  $q \in \Delta^*(s)$ , is a subtree of a tree  $s'$  satisfying  $p \in \Delta^*(s')$  which root is  $f_j$ , there exists a tree  $t$  in  $\mathcal{L}(\bar{\mathcal{C}}_E)$  for any tree  $s$  satisfying  $q \in \Delta^*(s)$  such that  $s \preceq t$ .

Suppose that  $q = C_{f_j^k}(\bar{E})$  is a coaccessible state in  $\bar{\mathcal{C}}_E$ . According to (**P'**), there exists a tree  $t$  in  $\mathcal{L}(\bar{\mathcal{C}}_E)$  for any tree  $s$  satisfying  $q \in \Delta^*(s)$  such that  $s \preceq t$ . By construction, since  $C_{f_j^k}(\bar{E})$  is coaccessible by the symbol  $f_j$ , there exists a tree  $s' \preceq t$  such that  $\text{root}(s') = f_j$  and  $k\text{-child}(s') = \text{root}(s)$ . By definition  $\text{root}(s) \in \text{Follow}(\bar{E}, f_j, k)$ . As a direct consequence, if  $C_{f_j^k}(\bar{E})$  is a coaccessible state in  $\mathcal{C}_E$ , it is in  $\bar{\mathcal{C}}_E$ . As previously shown, this implies that  $\text{Follow}(\bar{E}, f_j, k) \neq \emptyset$ . As a conclusion, by definition,  $\text{Follow}(\bar{E}, f_j, k) = \text{Follow}(E, f_j, k) \neq \emptyset$ .  $\square$

### 3.3 From $k$ -C-Continuation Automaton to Equation Automaton

The equation automaton is a quotient of the C-Continuation one w.r.t. the equivalence relation denoted by  $\sim_e$  over the set of states of  $\bar{\mathcal{C}}_E$  defined for any two states  $q_1 = C_{f_j^k}(\bar{E})$  and  $q_2 = C_{g_i^p}(\bar{E})$  by  $q_1 \sim_e q_2 \Leftrightarrow h(q_1) = h(q_2)$ .

**Proposition 5.** *The coaccessible part of the finite tree automaton  $\mathcal{C}_E / \sim_e$  is isomorphic to the equation tree automaton  $\mathcal{A}_E$ .*

*Proof.* Let  $E$  be a regular expression over an alphabet  $\Sigma$ . We define the inverse function of  $h$  denoted by  $h^{-1} : \Sigma \rightarrow 2^{\text{Pos}_E(E)}$  such that for any symbol  $x$  in  $\Sigma$ ,  $h^{-1}(x) = \{y \in \text{Pos}_E(E) \mid h(y) = x\}$ .

**Theorem 2 ([8]).** *Let  $E$  be a regular expression over an alphabet  $\Sigma$ . Then for every  $u \in \Sigma_{\geq 1}^*$ ,  $\partial_u(E) = \bigcup_{\bar{u} \in h^{-1}(u)} h(\partial_{\bar{u}}(\bar{E}))$ .*

**Proposition 6 ([8]).** *Let  $E$  be a regular expression over a ranked alphabet  $\Sigma$ . Then we have for every  $f \in \Sigma_{\geq 1}$ ,*

$$f^{-1}(E) = \bigcup_{f_j \in h^{-1}(f)} h(f_j^{-1}(\bar{E})) \text{ and } \partial_f(E) = \bigcup_{f_j \in h^{-1}(f)} h(\partial_{f_j}(\bar{E}))$$

Let us denote by  $\mathfrak{C}$  the coaccessible part of the finite tree automaton  $\mathcal{C}_E / \sim_e$ .

Let  $Q$  be the set of states of  $\mathcal{A}_E$  and  $Q'$  be the set of states of  $\mathfrak{C}$ . The isomorphism of the sets of states can be shown by the function  $\phi : Q' \rightarrow Q : [C_{f_j^k}(E)] \mapsto h(C_{f_j^k}(E))$ . Indeed, according to Lemma 2,  $C_{f_j^k}(E) \in \partial_{\bar{u}}(\bar{E})$  for some  $\bar{u} \in \text{Pos}_E(E)^*_{\geq 1}$ . Using Theorem 2,  $h(C_{f_j^k}(E)) = h(\partial_{\bar{u}}(\bar{E})) \in \partial_{\Sigma^*_{\geq 1}}(E) = Q$ . Injectivity of  $\phi$  can be shown directly from the definition of the equivalence relation  $\sim_e$ . For surjectivity, it is deduced from the Theorem 2.

By definition,  $\phi([C_{\varepsilon^1}(E)]) = C_{\varepsilon^1}(E) = E$ . Hence the image of the final state of  $\mathfrak{C}$  is the final state of  $\mathcal{A}_E$ .

Let us show that the sets of transitions are also isomorphic.

Let  $([C_{f_j^k}(E)], g, [C_{g_i^1}(E)] \dots, [C_{g_i^m}(E)])$  be a transition in  $\mathfrak{C}$ . Equivalently by construction, there exists a symbol  $g_i$  such that  $(C_{f_j^k}(E), g_i, C_{g_i^1}(E) \dots, C_{g_i^m}(E))$  is a transition in the accessible part of the automaton  $\bar{\mathcal{C}}_E$ . As the coaccessible part of  $\bar{\mathcal{C}}_E$  and  $\mathcal{A}_{\bar{E}}$  are equal (by Lemma 2), the transition  $(\bar{F}, g_i, \bar{H}_1, \dots, \bar{H}_m)$  is in the automaton  $\mathcal{A}_{\bar{E}}$  with  $\bar{F} = C_{f_j^k}(E)$  and  $\bar{H}_l = C_{g_i^l}(E)$  for  $1 \leq l \leq m$ ; consequently  $(\bar{H}_1, \dots, \bar{H}_m) \in g_i^{-1}(\bar{F})$ . From Proposition 6, it is equivalent to  $(H_1 \dots, H_m) \in g^{-1}(F)$ . Thus  $(F, g, H_1 \dots, H_m) = (h(C_{f_j^k}(E)), g, h(C_{g_i^1}(E)), \dots, h(C_{g_i^m}(E)))$  is a transition in the automaton  $\mathcal{A}_E$ .

Since only equivalences are stated,  $([C_{f_j^k}(E)], g, [C_{g_i^1}(E)] \dots, [C_{g_i^m}(E)])$  is a transition in  $\mathfrak{C}$  if and only if  $(\phi([C_{f_j^k}(E)]), g, \phi([C_{g_i^1}(E)]), \dots, \phi([C_{g_i^m}(E)]))$  is a transition in  $\mathcal{A}_E$ .

Finally, for  $c \in \Sigma_0$ ,  $([C_{f_j^k}(E)], c)$  is a transition in  $\mathfrak{C}$  if and only if  $c \in \llbracket h(C_{f_j^k}(E)) \rrbracket$ . Furthermore, it holds by construction that  $(\phi([C_{f_j^k}(E)]), c)$  is a transition in  $\mathcal{A}_E$  if and only if  $c \in \llbracket \phi([C_{f_j^k}(E)]) \rrbracket$ . Consequently,  $([C_{f_j^k}(E)], c)$  is a transition in  $\mathfrak{C}$  if and only if  $(\phi([C_{f_j^k}(E)]), c)$  is a transition in  $\mathcal{A}_E$ .

As a conclusion,  $\phi$  is an isomorphism between  $\mathfrak{C}$  and  $\mathcal{A}_E$ .  $\square$

## 4 Construction of the equation tree automaton $\mathcal{A}_E$

In [8], the computation of the  $k$ -C-Continuations requires a preprocessing step which is the identification of subexpression of  $E$  in  $O(|E|^2)$  time and space complexity. We propose an algorithm for the computation of the set of states with an  $O(|E|)$  time and space complexity.

### 4.1 Computation of the set of states $Q_{\bar{\mathcal{C}}} / \sim_e$

The main idea is to efficiently compute the quotient  $\bar{\mathcal{C}}_E / \sim_e$  by converting the syntax tree into a finite acyclic deterministic word automaton.

Let  $T_E$  be the syntax tree associated with  $E$ . The set of nodes of  $T_E$  is written as  $\text{Nodes}(E)$ . For a node  $\nu$  in  $\text{Nodes}(E)$ ,  $\text{sym}(\nu)$ ,  $\text{father}(\nu)$ ,  $\text{son}(\nu)$ ,  $\text{right}(\nu)$  and

$\text{left}(\nu)$  denote respectively the symbol, the father, the son, the right son and the left son of the node  $\nu$  if they exist. We denote by  $E_\nu$  the subexpression rooted at  $\nu$ ; In this case we write  $\nu_E$  to denote the node associated to  $E_\nu$ . Let  $\gamma : \text{Nodes}(E) \cup \{\perp\} \rightarrow \text{Nodes}(E) \cup \{\perp\}$  be the function defined by:

$$\gamma(\nu) = \begin{cases} \text{father}(\nu) & \text{if } \text{sym}(\text{father}(\nu)) = {}^*c \text{ and } \nu \neq \nu_E \\ \text{right}(\text{father}(\nu)) & \text{if } \text{sym}(\text{father}(\nu)) = \cdot c \\ \perp & \text{otherwise} \end{cases}$$

where  $\perp$  is an artificial node such that  $\gamma(\perp) = \perp$ . The ZPC-Structure is the syntax tree equipped with  $\gamma(\nu)$  links. We extend the relation  $\preccurlyeq$  to the set of nodes of  $T_E$ : For two nodes  $\mu$  and  $\nu$  we write  $\nu \preccurlyeq \mu \Leftrightarrow T_{E_\nu} \preccurlyeq T_{E_\mu}$ . We define the set  $\Gamma_\nu(E) = \{\mu \in \text{Nodes}(E) \mid \nu \preccurlyeq \mu \wedge \gamma(\mu) \neq \perp\}$  which is totally ordered by  $\preccurlyeq$ .

**Proposition 7.** *Let  $E$  be linear,  $1 \leq k \leq n$  be two integers and  $f$  be in  $\Sigma_E \cap \Sigma_n$ . Then  $C_{f^k}(E) = (((E_{\nu_0} \cdot \text{op}(\nu_1) E_{\gamma(\nu_1)}) \cdot \text{op}(\nu_2) E_{\gamma(\nu_2)}) \cdots \text{op}(\nu_m) E_{\gamma(\nu_m)})$  where  $\nu_f$  is the node of  $T_E$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot c, {}^*c\}$ .*

*Proof.* By induction over the structure of  $E$ .

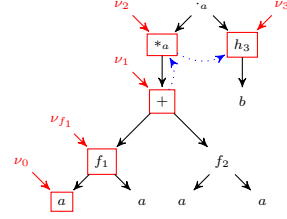
1. Let us suppose that  $E = f(E_1, \dots, E_n)$ . Then  $C_{f^k}(E) = E_k$ . Since by definition  $\nu_f$  is the root of  $T_E$ ,  $k$ -child( $\nu_f$ ) is the root of  $E_k$ . Hence  $E_{\nu_0} = E_k = C_{f^k}(E)$ .
2. Let us suppose that  $E = g(E_1, \dots, E_m)$  with  $g \neq f$ , or  $E = E_1 + E_2$ , or  $E = E_1 \cdot c E_2$  with  $f \in \Sigma_{E_2}$ . Then  $C_{f^k}(E) = C_{f^k}(E_j)$  with  $f \in \Sigma_{E_j}$ . By induction hypothesis,  $C_{f^k}(E_j) = (((E_{\nu_0} \cdot \text{op}(\nu_1) E_{\gamma(\nu_1)}) \cdot \text{op}(\nu_2) E_{\gamma(\nu_2)}) \cdots \text{op}(\nu_m) E_{\gamma(\nu_m)})$  where  $\nu_f$  is the node of  $T_{E_j}$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E_j) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot c, {}^*c\}$ . Since  $T_{E_j} \preccurlyeq T_E$ ,  $C_{f^k}(E_j) = (((E_{\nu_0} \cdot \text{op}(\nu_1) E_{\gamma(\nu_1)}) \cdot \text{op}(\nu_2) E_{\gamma(\nu_2)}) \cdots \text{op}(\nu_m) E_{\gamma(\nu_m)})$  where  $\nu_f$  is the node of  $T_E$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E_j) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot c, {}^*c\}$ .
3. Let us suppose that  $E = E_1 \cdot c E_2$  with  $f \in \Sigma_{E_1}$  (resp.  $E = E_1 {}^*c$ ). Then  $C_{f^k}(E) = C_{f^k}(E_1) \cdot c G$  with  $G \in \{E_1 {}^*c, E_2\}$ . By induction hypothesis,  $C_{f^k}(E_1) = (((E_{\nu_0} \cdot \text{op}(\nu_1) E_{\gamma(\nu_1)}) \cdot \text{op}(\nu_2) E_{\gamma(\nu_2)}) \cdots \text{op}(\nu_m) E_{\gamma(\nu_m)})$  where  $\nu_f$  is the node of  $T_{E_1}$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E_1) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot c, {}^*c\}$ . Since  $T_{E_1} \preccurlyeq T_E$ , by setting  $H = E_{\nu_{m+1}}$  and  $\text{op}(\nu_{m+1})c$ ,  $C_{f^k}(E_1) \cdot c H = (((E_{\nu_0} \cdot \text{op}(\nu_1) E_{\gamma(\nu_1)}) \cdot \text{op}(\nu_2) E_{\gamma(\nu_2)}) \cdots \text{op}(\nu_m) E_{\gamma(\nu_m)}) \cdot \text{op}(\nu_{m+1}) E_{\gamma(\nu_{m+1})}$  where  $\nu_f$  is the node of  $T_E$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E) = \{\nu_1, \dots, \nu_m, \nu_{m+1}\}$  and for  $1 \leq i \leq m+1$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot c, {}^*c\}$ .

□

**Corollary 3.** *Let  $E$  be linear,  $f \in (\Sigma_E)_m$  and  $k \leq m$ . Then  $|C_{f^k}(E)| \leq |E|^2$ .*

*Example 2.* Let  $\Sigma$  be the ranked alphabet such that  $\Sigma_0 = \{a, b\}$ ,  $\Sigma_1 = \{h\}$  and  $\Sigma_2 = \{f\}$ . Let  $E = (f(a, a) + f(a, a)) {}^*a \cdot a h(b)$ . Then  $\bar{E} = (f_1(a, a) + f_2(a, a)) {}^*a \cdot a h_3(b)$ . The ZPC-Structure associated with  $\bar{E}$  is represented in Figure 1 restricted to some  $\gamma$  links. As stated in Proposition 7,  $C_{f_1^1}(\bar{E}) = ((a \cdot a (f_1(a, a) + f_2(a, a)) {}^*a) \cdot a h_3(b)) = ((E_{\nu_0} \cdot a E_{\gamma(\nu_1)}) \cdot a E_{\gamma(\nu_2)})$ .

In order to identify the equivalent  $k$ -C-Continuations, we can sort them in lexicographic order. This can be done in  $O(|E|^3)$  time and space complexity using Paige and Tarjan's Algorithm [12]. This is due to the fact that the size of  $k$ -C-Continuations is in  $O(|E|^2)$  (by Corollary 3). This complexity has been improved by using  $k$ -Pseudo-Continuations instead of  $k$ -C-Continuations [3,7].



**Fig. 1.** ZPC-Structure of  $E$ .

A  $k$ -Pseudo-Continuation  $l_{f_j^k}(E)$  of  $f_j$  in  $E$  is obtained from the  $k$ -C-Continuation  $C_{f_j^k}(\bar{E})$  by replacing some subexpression  $\bar{F}$  of  $\bar{E}$  by a symbol  $\psi(h(\bar{F}))$  such that for two subexpressions  $F$  and  $G$  of  $E$ :  $\psi(F) = \psi(G) \Leftrightarrow F = G$ .

**Definition 5.** Let  $H$  be a regular expression over  $\Sigma$  and  $\psi$  be a bijection that associates to each subexpression of  $E$  a symbol in an alphabet  $\Psi$ . We define the word  $\psi'(H)$  over the alphabet  $\Psi \cup \{\cdot_a \mid a \in \Sigma_0\}$  inductively as follows:

$$\psi'(H) = \begin{cases} \psi'(F) \cdot_c \psi'(G) & \text{if } H = F \cdot_c G \text{ and } G \text{ a subexpression of } E \\ \psi(H) & \text{if } H \neq F \cdot_c G \text{ and } H \text{ a subexpression of } E \\ \varepsilon & \text{otherwise.} \end{cases}$$

The function  $\psi'$  is said to be an  $(E, \Psi)$ -encoding.

**Definition 6.** Let  $n$  and  $k$  be two integers such that  $1 \leq k \leq n$ ,  $f_j$  be a symbol in  $\text{Pos}_E(E)$  and  $\psi'$  an  $(E, \Psi)$ -encoding for some alphabet  $\Psi$ . The  $k$ -Pseudo-Continuation of  $f_j$  in  $E$ , denoted by  $l_{f_j^k}(E)$ , is the word over  $\Psi \cup \{\cdot_a \mid a \in \Sigma_0\}$  defined by  $l_{f_j^k}(E) = \psi'(h(C_{f_j^k}(\bar{E})))$ .

In the following, we consider that the pseudo-continuations of  $E$  are defined over  $\Psi$  a finite subset of  $\mathbb{N}$ , bounded by the number of subexpressions of  $E$ .

**Lemma 4.** Let  $E$  and  $F$  be two regular expressions over an alphabet  $\Sigma$  such that  $E$  and  $F$  are two products of subexpressions of a regular expression  $H$  over  $\Sigma$ . Let  $\psi'$  be a  $(H, \Psi)$ -encoding. Then:

$$\psi'(E) = \psi'(F) \Leftrightarrow E = F.$$

*Proof.* Let us consider that  $\psi'$  is associated with the bijection  $\psi$ . Let us consider the possible roots of the expressions.

1. If the roots of  $E$  and  $F$  are notconcatenation products,  $\psi'(E) = \psi'(F) \Leftrightarrow E = F$  since  $\psi$  is a bijection and  $\psi'(E) = \psi(E)$  wedge  $\psi'(F) = \psi(F)$ .
2. Let us suppose that, without loss of generality, only the root of  $E$  is a concatenation product  $\cdot_c$ . Then the symbol  $\cdot_c$  appears in  $\psi'(E)$  but not in  $\psi'(F)$ . Hence  $\psi'(E) \neq \psi'(F)$  and  $E \neq F$ .
3. Finally, let us suppose that  $E = E_1 \cdot_c E_2$  and  $F = F_1 \cdot_d F_2$ .
  - (a) If  $E_2 \neq F_2$  then  $\psi(E) \neq \psi(F)$  and then  $\psi'(E)$  and  $\psi'(F)$  do not end with the same symbol.

- (b) Suppose that  $E_2 = F_2$ . If  $\cdot_c \neq \cdot_d$ ,  $\psi'(E)$  ends with  $\cdot_c \psi(E_2) \neq \cdot_d \psi(E_2)$ ; Hence  $\psi'(E) \neq \psi'(F)$  and  $E \neq F$ . Otherwise, by induction hypothesis,  $\psi'(E_1) = \psi'(F_1) \Leftrightarrow E_1 = F_1$ . Hence  $\psi'(E_1) \cdot_c \psi(E_2) = \psi'(F_1) \cdot_c \psi(F_2) \Leftrightarrow E_1 \cdot_c E_2 = F_1 \cdot_d F_2$ .

□

**Proposition 8.** *The two following propositions hold:*

1.  $|l_{f_j^k}(E)|$  is at most linear w.r.t.  $|E|$ ,
2.  $\sum_{f_j \in \text{Pos}_E(E)_n, 1 \leq k \leq n} |\psi'(E_{\nu_{f_j^k}})|$  is at most linear w.r.t.  $|E|$ , with  $\nu_{f_j^k} = k\text{-child}(\nu_{f_j})$ .

*Proof.* We define the function  $\text{nbdot}(E)$  as the number of left-associated concatenation operators in  $E$  as follows:

$$\text{nbdot}(E) = \begin{cases} \text{nbdot}(F) + 1 & \text{if } E = F \cdot_c G, \\ 0 & \text{otherwise.} \end{cases}$$

Let us first prove that  $|\psi'(E)| \leq 2\text{nbdot}(E) + 1$ . The proof proceeds by induction in the structure of  $E$ .

1. Whenever  $E$  is not a product,  $\psi'(E) = \psi(E)$ . Hence  $|\psi'(E)| = 1$ . Since  $\text{nbdot}(E) = 0$ , the condition is satisfied.
2. Suppose that  $E = F \cdot_c G$ . Hence

$$\begin{aligned} |\psi'(F \cdot_c G)| &\leq |\psi'(F)| + 2 \\ &\leq 2(\text{nbdot}(F)) + 1 + 2 \\ &= 2(\text{nbdot}(F) + 1) + 1 \\ &= 2(\text{nbdot}(E)) + 1. \end{aligned}$$

Following Proposition 7,  $|l_{f_j^k}(E)| \leq |\psi'(E_{\nu_0})| + 2m$  where  $\nu_f$  is the node of  $T_E$  labelled by  $f$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_f}(E) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot_c, \cdot_e\}$ . Since  $\{\nu_1, \dots, \nu_m\}$  are ancestors of  $\nu_0$ ,  $|E_{\nu_0}| \leq |E| - m$ . Consequently,  $\text{nbdot}(E_{\nu_0}) \leq |E| - m$ . Moreover, from previous point,  $|\psi'(E_{\nu_0})| \leq 2(\text{nbdot}(E_{\nu_0})) + 1$ . Consequently,  $|l_{f_j^k}(E)| \leq 2(|E| - m) + 1 + 2m = 2|E| + 1$ .

Furthermore, since  $|\psi'(E_{\nu_{f_j^k}})| \leq 2(\text{nbdot}(E_{\nu_{f_j^k}})) + 1$ , it holds:

$$\sum_{f_j \in \text{Pos}_E(E)_n} \sum_{1 \leq k \leq n} |\psi'(E_{\nu_{f_j^k}})| \leq \sum_{f_j \in \text{Pos}_E(E)_n} \sum_{1 \leq k \leq n} 2(\text{nbdot}(E_{\nu_{f_j^k}})) + 1.$$

However, the concatenation operators below the node  $\nu_{f_j^k}$  do not appears below another symbol. Consequently,

$$\sum_{f_j \in \text{Pos}_E(E)_n} \sum_{1 \leq k \leq n} \text{nbdot}(E_{\nu_{f_j^k}}) \leq |E|$$

Finally,

$$\sum_{f_j \in \text{Pos}_E(E)_n} \sum_{1 \leq k \leq n} |\psi'(E_{\nu_{f_j^k}})| \leq 2|E| + |\Sigma_{\geq 1}|.$$

□

**Proposition 9.** *Let  $f_j \in \text{Pos}_E(E)_n$ ,  $g_i \in \text{Pos}_E(E)_m$ ,  $k \leq n$  and  $p \leq m$  be two integers. Then  $h(C_{f_j^k}(\bar{E})) = h(C_{g_i^p}(\bar{E})) \Leftrightarrow l_{f_j^k}(E) = l_{g_i^p}(E)$ .*

*Proof.* Direct Corollary of Lemma 4.

□

From Proposition 9 we can deduce that the  $k$ -C-Continuations identification can be achieved by considering the  $k$ -Pseudo-Continuations. In the following we show that this identification step (computation of  $\sim_e$ ) can be done without the computation of the  $k$ -Pseudo-Continuations and that it amounts to the minimization of a word acyclic deterministic automaton. Before seeing how the identification of  $k$ -Pseudo-Continuations  $l_{f_j^k}(E)$  is performed, we prove that the computation of the function  $\psi$  can be done in a linear time in the size of  $E$ .

Let us consider the syntax tree  $T_E$  associated with  $E$ . This syntax tree contains all the subexpressions of  $E$ . Each node  $\nu$  in  $T_E$  corresponds to the subexpression  $E_\nu$  of  $E$ . The equivalence relation  $\sim$  over the nodes of the tree  $T_E$  is defined by  $\nu_1 \sim \nu_2 \Leftrightarrow E_{\nu_1} = E_{\nu_2}$ . We show that the computation of the equivalence relation  $\sim$  amounts to the minimization of the word acyclic deterministic automaton  $\mathcal{A}_{T_E} = (Q, \Sigma_A, \{\nu_E\}, \{\nu_T\}, \delta)$ , where  $\nu_E$  is the node associated to the root of  $E$ ,  $Q = \text{Nodes}(E) \cup \{\nu_T\} \cup \{\perp\}$  with  $\nu_T, \perp \notin \text{Nodes}(E)$ ,  $\Sigma_A = \Sigma_0 \cup \{g_+, d_+\} \cup \{*_a, g_{\cdot a}, d_{\cdot a} \mid a \in \Sigma_0\} \cup \{f^1, \dots, f^n \mid f \in \Sigma_n, n \geq 1\}$ , and  $\delta$  is defined by  $\delta(\nu, *_a) = \text{son}(\nu)$  if  $\text{sym}(\nu) = *_a$ ,  $\delta(\nu, g_{\text{sym}(\nu)}) = \text{left}(\nu)$  and  $\delta(\nu, d_{\text{sym}(\nu)}) = \text{right}(\nu)$  if  $\text{sym}(\nu) \in \{+, \cdot_a, a \in \Sigma_0\}$ ,  $\delta(\nu, \text{sym}(\nu)) = \nu_T$  if  $\text{sym}(\nu) \in \Sigma_0$ ,  $\delta(\nu, f^k) = k\text{-child}(\nu)$  if  $\text{sym}(\nu) = f \in \Sigma_{\geq 1}$ , and  $\delta(\nu, x) = \perp$  in all otherwise.

**Lemma 5.**  $E = F \Leftrightarrow \mathcal{L}(\mathcal{A}_{T_E}) = \mathcal{L}(\mathcal{A}_{T_F})$ .

*Proof.* Let  $\Sigma_{\mathcal{A}_E}$  (resp.  $\Sigma_{\mathcal{A}_F}$ ) be the alphabet of the automaton  $\mathcal{A}_{T_E}$  (resp.  $\mathcal{A}_{T_F}$ ).

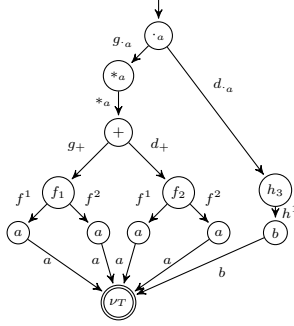
1. If  $E = F$  then  $\mathcal{A}_{T_E} = \mathcal{A}_{T_F}$  and  $\mathcal{L}(\mathcal{A}_{T_E}) = \mathcal{L}(\mathcal{A}_{T_F})$ .
2. Suppose that  $E \neq F$ . Notice that any word  $w$  in  $\mathcal{L}(\mathcal{A}_{T_E})$  (resp.  $\mathcal{L}(\mathcal{A}_{T_F})$ ) starts with a symbol associated with the root of  $E$  (resp.  $F$ ).
  - (a) Hence, if the roots of  $E$  and  $F$  are distinct, then  $\mathcal{L}(\mathcal{A}_{T_E}) \cap \mathcal{L}(\mathcal{A}_{T_F}) = \emptyset$ ; Since  $\mathcal{L}(\mathcal{A}_{T_E})$  is not empty by construction,  $\mathcal{L}(\mathcal{A}_{T_E}) \neq \mathcal{L}(\mathcal{A}_{T_F})$ .
  - (b) Otherwise, there exists an integer  $j$  such that  $E = x(E_1, \dots, E_k)$ ,  $F = x(F_1, \dots, F_k)$  and  $E_j \neq F_j$ . By induction over the size of  $E_j$ .
    - i. If  $E_j \in \Sigma_0$ , then since the roots are distincts, the word starting with the symbol associated to the node  $x$  followed by the symbol  $a$  is in  $\mathcal{L}(\mathcal{A}_{T_E})$  but not in  $\mathcal{L}(\mathcal{A}_{T_F})$ .
    - ii. Otherwise, it holds by induction hypothesis that there exists a word in  $\mathcal{L}(\mathcal{A}_{T_{E_j}})$  not in  $\mathcal{L}(\mathcal{A}_{T_{F_j}})$ . Hence there exists a word starting with a symbol associated to the node  $x$  followed by a word in  $\mathcal{L}(\mathcal{A}_{T_{E_j}})$  that is in  $\mathcal{L}(\mathcal{A}_{T_E})$  but not in  $\mathcal{L}(\mathcal{A}_{T_F})$ .

□

According to Lemma 5,  $\nu_1 \sim \nu_2 \Leftrightarrow \mathcal{L}(\mathcal{A}_{T_{E_{\nu_1}}}) = \mathcal{L}(\mathcal{A}_{T_{E_{\nu_2}}})$ , that is the equivalence relation  $\sim$  coincides with Myhill-Nerode equivalence [11] over the states of the automaton  $\mathcal{A}_{T_E}$ , that can be computed in  $O(|E|)$  time and space complexity using Revuz Algorithm [13].

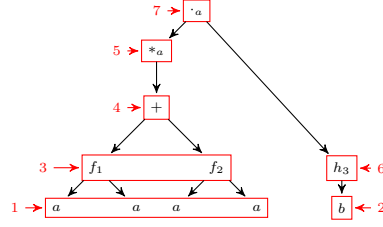
**Lemma 6.** *The computation of  $\psi(F)$  for all subexpression  $F$  of  $E$  can be done in  $O(|E|)$  time and space complexity.*

*Proof.* Let  $\nu_1$  and  $\nu_2$  be two nodes in  $T_E$ . As  $\nu_1 \sim \nu_2 \Leftrightarrow E_{\nu_1} = E_{\nu_2} \Leftrightarrow \psi(E_{\nu_1}) = \psi(E_{\nu_2})$ , we can associate to each node  $\nu$  in  $T_E$  (each  $E_\nu$ ) a symbol ( $\psi(E_\nu)$ ) which uniquely identifies its equivalence class  $[\nu]_\sim$ . Furthermore, according to Lemma 5, the computation of the equivalence relation  $\sim$  amounts to the minimization of the word acyclic deterministic automaton  $\mathcal{A}_{T_E}$ , which can be performed in  $O(|E|)$  using Revuz Algorithm [13].  $\square$



**Fig. 2.** The automaton  $\mathcal{A}_{T_E}$ .

*Example 3.* Let us consider the regular expression  $E = (f(a, a) + f(a, a))^* \cdot a \cdot h(b)$  of the Example 2. Applying Myhill-Nerode equivalence [11] to the states of the automaton  $\mathcal{A}_{T_E}$  (Figure 2) results in 7 equivalence classes labeled by  $\Psi = \{1, 2, \dots, 7\}$ . For example  $\psi(f(a, a)) = 3$  and  $\psi(E) = 7$  (Figure 3). Finally,  $l_{f_1^1}(E) = 1 \cdot a \cdot 6 \cdot a \cdot 5$ .



**Fig. 3.** The Equivalence Classes.

Recall that the  $k$ -Pseudo-Continuation identification can be achieved in  $O(|E|^2)$  [4,8] using Paige and Tarjan's sorting algorithm [12]. In what follows we show that this step amounts to the minimization of the acyclic deterministic word automaton  $\mathcal{B}_{T_E} = (Q_B, \Sigma_B, \{\nu_T\}, \{\nu_{\bar{E}}\}, \delta_B)$  defined with  $\nu_T \notin \text{Nodes}(\bar{E})$  and  $\mathfrak{F} = \{f_j^k \mid 1 \leq k \leq m, f_j \in \text{Pos}_E(E)_m\}$  by:

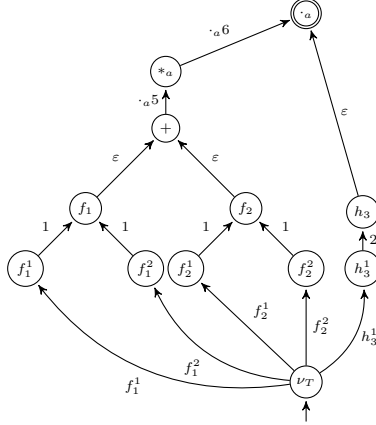
- $Q_B = (\text{Nodes}(\bar{E}) \setminus \Sigma_0) \cup \mathfrak{F} \cup \{\nu_T, \perp\}$ ,
- $\Sigma_B = \{\psi(\nu) \mid \nu \in \text{Nodes}(\bar{E}) \cap Q_B\} \cup \mathfrak{F} \cup \{\cdot_a \mid a \in \Sigma_0\} \cup \{\varepsilon\}$ ,
- $\delta_B$  is defined as follows:
  - $\delta(\nu_T, f_j^i) = f_j^i$  for all  $f_j^i \in \mathfrak{F}$ ,
  - $\delta(f_j^k, \psi'(h(E_{\nu_k}))) = f_j$  if  $\nu_k$  is the  $k^{\text{th}}$  child of  $f_j$ ,
  - $\delta(\nu, \cdot_a \psi(E_{\gamma(\nu)})) = \text{father}(\nu)$  if  $\text{sym}(\text{father}(\nu)) \in \{\cdot_a, *a\}$  and  $\gamma(\nu) \neq \perp$ ,
  - $\delta(\nu, \varepsilon) = \text{father}(\nu)$  and if  $\gamma(\nu) = \perp$  and  $\delta(\nu, x) = \perp$  in all otherwise.

**Proposition 10.**  $\mathcal{L}(\mathcal{B}_{T_E}) = \{f_j^k \cdot l_{f_j^k}(E) \mid f_j \in \text{Pos}_E(E)_m, k \leq m\}$

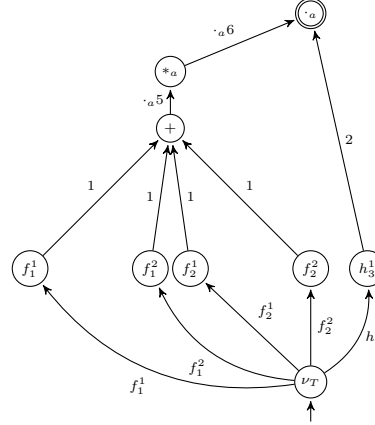
*Proof.* By construction of  $\mathcal{B}_{T_E}$ , there exists a path from any state  $f_j^k$  with  $f_j \in \text{Pos}_E(E)_m$  and  $1 \leq k \leq m$  to the root of  $E$  labelled by  $\psi'(E_{\nu_0}) \cdot \text{op}(\nu_1) \psi(E_{\gamma(\nu_1)}) \cdots \text{op}(\nu_m) \psi(E_{\gamma(\nu_m)})$  where  $\nu_{f_j}$  is the node of  $T_E$  labelled by  $f_j$ ,  $\nu_0$  is the  $k$ -child( $\nu_f$ ),  $\Gamma_{\nu_{f_j}}(E) = \{\nu_1, \dots, \nu_m\}$  and for  $1 \leq i \leq m$ ,  $\text{op}(\nu_i) = c$  such that  $\text{sym}(\text{father}(\nu_i)) \in \{\cdot_c, *c\}$ . This word exactly corresponds to the word  $\psi'(h(C_{f_j^k}(\bar{E}))) = l_{f_j^k}(E)$ .  $\square$



Let  $f_j$  and  $g_i$  be two positions in  $\text{Pos}_{\mathbb{E}}(\mathbb{E})$ . As a direct consequence of Proposition 10,  $C_{f_j^k}(\overline{\mathbb{E}}) \sim_e C_{g_i^p}(\overline{\mathbb{E}})$  if and only if the states  $f_j^k$  and  $g_i^p$  of  $\mathcal{B}_{T_{\overline{\mathbb{E}}}}$  are equivalent. We eliminate the  $\varepsilon$ -transitions from the automaton  $\mathcal{B}_{T_{\overline{\mathbb{E}}}}$ . Since it has no  $\varepsilon$ -transitions cycles, this elimination can be performed in a linear time in the size of  $\mathbb{E}$ . Hence, we obtain a more compacted but equivalent structure, which we denote by  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{\mathbb{E}}}})$ .



**Fig. 4.** The automaton  $\mathcal{B}_{T_{\overline{\mathbb{E}}}}$ .



**Fig. 5.** The automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{\mathbb{E}}}})$ .

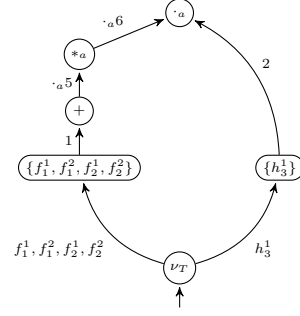
The computation of the equivalence relation  $\sim_e$  can be performed by the computation of Myhill-Nerode relation [11] on the states of the automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{\mathbb{E}}}})$ . This automaton is deterministic and acyclic.

**Theorem 3.** *The relation  $\sim_e$  can be computed in  $O(|\mathbb{E}|)$  time complexity.*

*Proof.* The equivalence relation  $\sim_e$  coincides with Myhill-Nerode equivalence [11] on the states of the automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{\mathbb{E}}}})$ .

This automaton is deterministic and acyclic and its size is linear with respect to  $|\mathbb{E}|$  (Proposition 8). That can be computed in  $O(|\mathbb{E}|)$  time and space complexity using Revuz Algorithm [13].  $\square$

*Example 4.* Let us consider the regular expression  $E = (f(a, a) + f(a, a))^* \cdot_a h(b)$  of Example 2. The automaton  $\mathcal{B}_{T_{\overline{E}}}$  is represented by Figure 4. The automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{E}}})$  is represented in Figure 5. Applying Myhill-Nerode equivalence to the automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{E}}})$  results in the automaton in Figure 6. We deduce from this automaton that  $C_{f_1^1}(\overline{E}) \sim_e C_{f_1^2}(\overline{E}) \sim_e C_{f_2^1}(\overline{E}) \sim_e C_{f_2^2}(\overline{E})$ . Consequently the set of states of  $\mathcal{C}_E/\sim_e$  is  $\{[C_{\varepsilon^1}(\overline{E})], [C_{f_1^1}(\overline{E})], [C_{h_3^1}(\overline{E})]\}$ .



**Fig. 6.** The Minimal Automaton of  $\varepsilon\text{-free}(\mathcal{B}_{T_{\overline{E}}})$ .

## 4.2 Computation of the set of transition rules

Using Proposition 3 and Proposition 4, we can show that the computation of the set of transitions of the equation tree automaton is performed by computing the function Follow. The computation of a transition rule using Proposition 3 requires a linear time, according to Proposition 2. Then for all transition rules we get an  $O(|Q/\sim_e| \times |E|)$  time and space complexity where  $Q$  is the set of  $k$ -C-Continuations of  $\overline{E}$ . The computation of the set of states  $Q_{\overline{E}}/\sim_e$  make possible the creation of non-coaccessible states. Removing these states requires an  $O(|Q_{\overline{E}}/\sim_e| \cdot |E|)$  time complexity.

**Theorem 4.** *The equation tree automaton  $\mathcal{A}_E$  of  $E$  can be computed in  $O(|Q| \cdot |E|)$  time and space complexity with  $Q$  the set of states of  $\mathcal{A}_E$ .*

*Proof.* The equivalence relation  $\sim_e$  can be computed in  $O(|E|)$  time and space complexity and the set of transition rules can be performed by computing the function Follow. The computation of a transition rule using Proposition 3 requires a linear time, according to Proposition 2. Then for all transition rules we get an  $O(|Q_{\overline{E}}/\sim_e| \times |E|)$  time and space complexity where  $Q_{\overline{E}}$  is the set of  $k$ -C-Continuations of  $\overline{E}$ . Finally, removing not coaccessible states can be performed in linear time and results in the equation automaton.  $\square$

## 5 A Full Example

Let  $E = h(h(c, b) \cdot_c a, a) \cdot_b (f(a, h(c, b)) \cdot_c a + g(a))^* \cdot_b h$  be a regular expression defined over the ranked alphabet  $\Sigma$  such that  $\Sigma^0 = \{a, b, c\}$ ,  $\Sigma^1 = \{g\}$ ,  $\Sigma^2 = \{f, h\}$  and  $\overline{E} = h_1(h_2(c, b) \cdot_c a, a) \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^* \cdot_b h$  be its linearized form.

The computation of the  $k$ -C-Continuations of the  $E$  using the Definition 3 is given in Table 1.

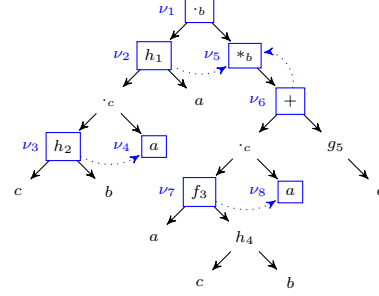
$$\begin{aligned}
C_{h_1^1}(\bar{E}) &= (h_2(c, b) \cdot_c a) \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{h_1^2}(\bar{E}) &= a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{f_3^1}(\bar{E}) &= a \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{f_3^2}(\bar{E}) &= (h_4(c, b) \cdot_c a) \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{h_2^1}(\bar{E}) &= c \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{h_2^2}(\bar{E}) &= b \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{h_4^1}(\bar{E}) &= c \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{h_4^2}(\bar{E}) &= b \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}, \\
C_{g_5^1}(\bar{E}) &= a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}.
\end{aligned}$$

**Table 1.** The  $k$ -C-Continuations of E.

From Table 1, the Follow function can be computed (Table 2).

$x_i^j$	$Follow(E, x_i, j)$
$h_1^1$	$\{h_2\}$
$h_1^2$	$\{a\}$
$h_2^1$	$\{a\}$
$h_2^2$	$\{g_5, f_3\}$
$f_3^1$	$\{a\}$
$f_3^2$	$\{h_4\}$
$h_4^1$	$\{a\}$
$h_4^2$	$\{g_5\}$
$g_5^1$	$\{a\}$

**Table 2.** The function Follow.



**Fig. 7.** The ZPC-Structure of E.

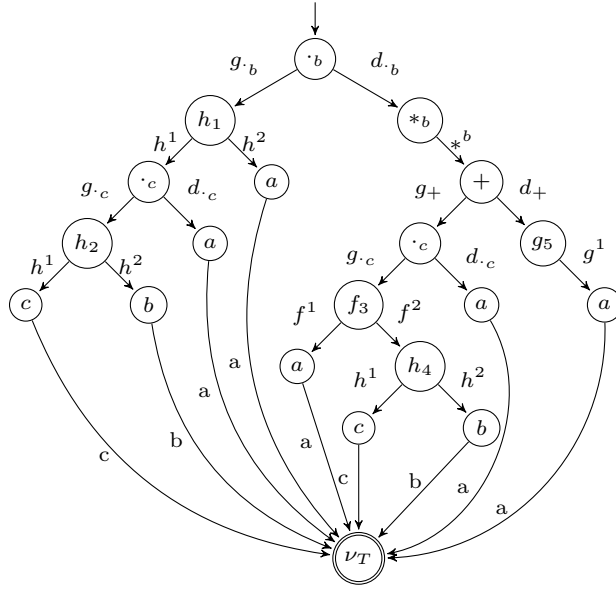
Finally, from Table 2, the transition function of  $\mathcal{C}_E$  is the following:

$$\begin{aligned}
h(C_{h_1^1}(\bar{E}), C_{h_1^2}(\bar{E})) &\rightarrow C_{\varepsilon^1}(\bar{E}) & h(C_{h_2^1}(\bar{E}), C_{h_2^2}(\bar{E})) &\rightarrow C_{h_1^1}(\bar{E}) \\
a &\rightarrow C_{h_1^2}(\bar{E}) & a &\rightarrow C_{h_2^1}(\bar{E}) \\
g(C_{g_5^1}(\bar{E})) &\rightarrow C_{h_2^2}(\bar{E}) & f(C_{f_3^1}(\bar{E}), C_{f_3^2}(\bar{E})) &\rightarrow C_{h_2^2}(\bar{E}) \\
a &\rightarrow C_{f_3^1}(\bar{E}) & h(C_{h_4^1}(\bar{E}), C_{h_4^2}(\bar{E})) &\rightarrow C_{f_3^2}(\bar{E}) \\
a &\rightarrow C_{g_5^1}(\bar{E}) & a &\rightarrow C_{h_4^1}(\bar{E}) \\
f(C_{f_3^1}(\bar{E}), C_{f_3^2}(\bar{E})) &\rightarrow C_{h_4^2}(\bar{E}) & g(C_{g_5^1}(\bar{E})) &\rightarrow C_{h_4^2}(\bar{E})
\end{aligned}$$

The ZPC-structure associated to E is represented in Figure 7. The dotted links in Figure 7 represent the function  $\gamma$ :

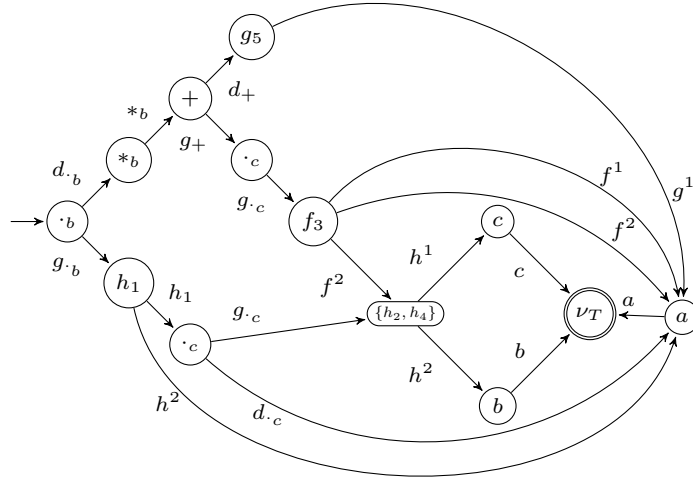
$$\gamma(\nu_2) = \nu_5, \gamma(\nu_3) = \nu_4, \gamma(\nu_6) = \nu_5, \gamma(\nu_7) = \nu_8.$$

The automaton  $\mathcal{A}_{T_E}$  associated with E is represented in Figure 8.



**Fig. 8.** The automaton  $\mathcal{A}_{T_E}$ .

Applying Myhill-Nerode equivalence relation over the states of the automaton  $\mathcal{A}_{T_E}$  results in the automaton in Figure 9.

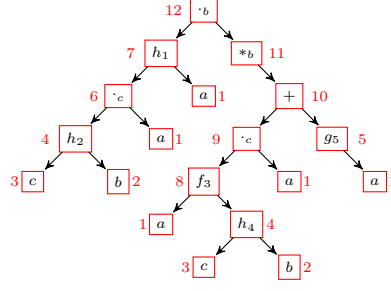


**Fig. 9.** The minimal automaton of  $\mathcal{A}_{T_E}$ .

The computation of the equivalence relation  $\sim$  over the syntax tree associated to E is represented in the Figure 10. The number of equivalence classes in Figure 10 (12) corresponds exactly to the number of states of the minimal automaton of  $\mathcal{A}_{T_E}$ . From these equivalence classes, we can define the  $\psi$  function (see Table 3).

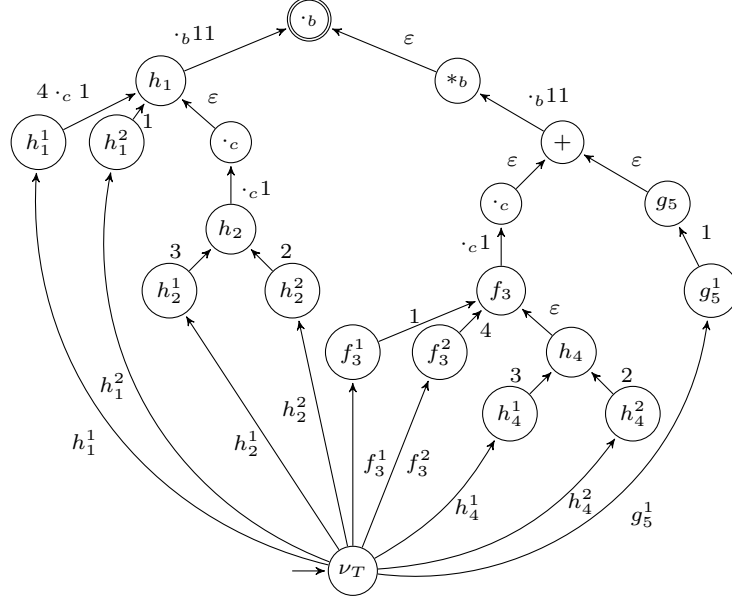
$$\begin{aligned}
 \psi(1) &= a \\
 \psi(2) &= b \\
 \psi(3) &= c \\
 \psi(4) &= h(c, b) \\
 \psi(5) &= g(a) \\
 \psi(6) &= h(c, b) \cdot_c a \\
 \psi(7) &= h(h(c, b) \cdot_c a, a) \\
 \psi(8) &= f(a, h(c, b)) \\
 \psi(9) &= f(a, h(c, b)) \cdot_c a \\
 \psi(10) &= f(a, h(c, b)) \cdot_c a + g(a) \\
 \psi(11) &= (f(a, h(c, b)) \cdot_c a + g(a)) * b \\
 \psi(12) &= E
 \end{aligned}$$

**Table 3.** The function  $\psi$ .



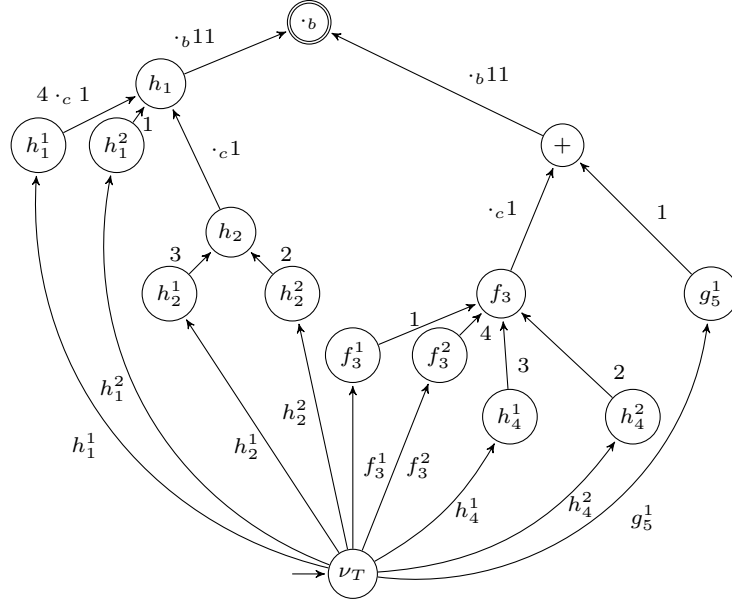
**Fig. 10.** The equivalence classes.

As we have seen, the computation of the equivalence relation  $\sim_e$  turns in the minimization of an acyclic deterministic automaton. The automaton  $\mathcal{B}_{T_E}$  associated with E is represented in Figure 11.



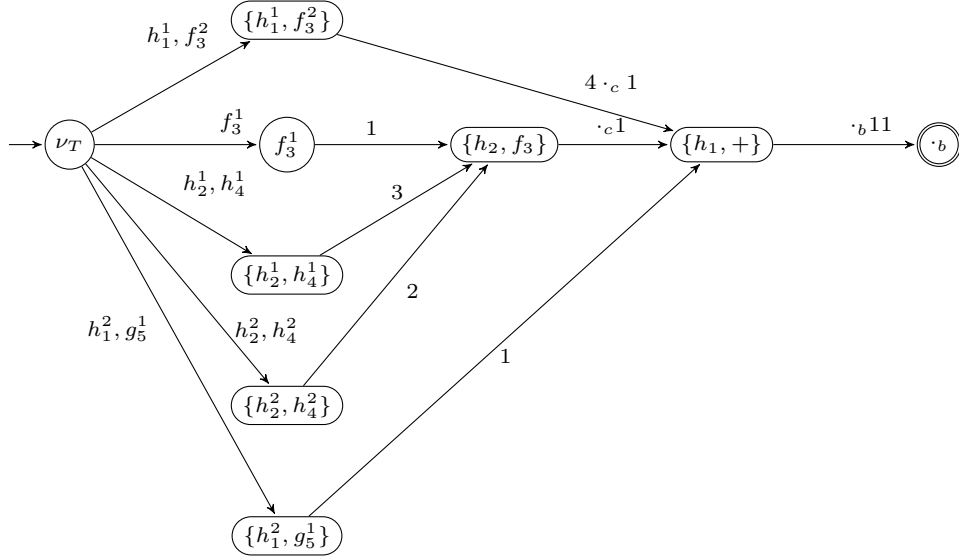
**Fig. 11.** The automaton  $\mathcal{B}_{T_E}$ .

We eliminate the  $\varepsilon$ -transitions from the automaton  $\mathcal{B}_{T_E}$ . Since this last has no  $\varepsilon$ -transitions cycles, this elimination can be performed in a linear time in the size of E. Hence, we obtain a structure which we denote  $\varepsilon\text{-free}(\mathcal{B}_{T_E})$ .



**Fig. 12.** The automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_E})$ .

The computation of the equivalence relation  $\sim_e$  amounts to apply Myhill-Nerode relation on the states of the automaton  $\varepsilon\text{-free}(\mathcal{B}_{T_E})$ . The result is represented in Figure 13.



**Fig. 13.** The Minimal Automaton of  $\varepsilon\text{-free}(\mathcal{B}_{T_E})$ .

The language recognized by  $\mathcal{B}_{T_E}$  is the following:

$$\begin{aligned}
\mathcal{L}(\mathcal{B}_{T_{\overline{E}}}) = & \{h_2^2, h_4^2\} \cdot \{2 \cdot_c 1 \cdot_b 11\} \\
& \cup \{h_2^1, h_4^1\} \cdot \{3 \cdot_c 1 \cdot_b 11\} \\
& \cup \{h_1^1, f_3^2\} \cdot \{4 \cdot_c 1 \cdot_b 11\} \\
& \cup \{h_1^2, g_5^1\} \cdot \{1 \cdot_b 11\} \\
& \cup \{f_3^1 1 \cdot_c 1 \cdot_b 11\}
\end{aligned}$$

Let us notice that Proposition 9 is satisfied in Table 4.

$x$	$xw \in \mathcal{L}(\mathcal{B}_{T_{\overline{E}}})$	$C_x(\overline{E})$
$h_1^1$	$h_1^1 4 \cdot_c 1 \cdot_b 11$	$(h_2(c, b) \cdot_c a) \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$h_1^2$	$h_1^2 1 \cdot_b 11$	$a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$h_2^1$	$h_2^1 3 \cdot_c 1 \cdot_b 11$	$c \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$h_2^2$	$h_2^2 2 \cdot_c 1 \cdot_b 11$	$b \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$f_3^1$	$f_3^1 1 \cdot_c 1 \cdot_b 11$	$a \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$f_3^2$	$f_3^2 4 \cdot_c 1 \cdot_b 11$	$(h_2(c, b) \cdot_c a) \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$h_4^1$	$h_4^1 3 \cdot_c 1 \cdot_b 11$	$c \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$h_4^2$	$h_4^2 2 \cdot_c 1 \cdot_b 11$	$b \cdot_c a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$
$g_5^1$	$g_5^1 1 \cdot_b 11$	$a \cdot_b (f_3(a, h_4(c, b)) \cdot_c a + g_5(a))^{*b}$

**Table 4.**  $\mathcal{L}(\mathcal{B}_{T_{\overline{E}}})$  and  $k$ -C-Continuations.

Finally, the equation automaton  $\mathcal{A}_{\overline{E}}$  associated with  $\overline{E}$  is obtained from merging the states and the transitions using  $\sim_e$ . The transition function is:

$$\begin{aligned}
h(\{C_{h_1^1}(\overline{E}), C_{f_3^2}(\overline{E})\}, \{C_{h_1^2}(\overline{E}), C_{g_5^1}(\overline{E})\}) & \rightarrow C_{\varepsilon^1}(\overline{E}) \\
& a \rightarrow \{C_{h_1^2}(\overline{E}), C_{g_5^1}(\overline{E})\} \\
h(\{C_{h_2^1}(\overline{E}), C_{h_4^1}(\overline{E})\}, \{C_{h_2^2}(\overline{E}), C_{h_4^2}(\overline{E})\}) & \rightarrow \{C_{h_1^1}(\overline{E}), C_{f_3^2}(\overline{E})\} \\
& a \rightarrow \{C_{h_2^1}(\overline{E}), C_{h_4^1}(\overline{E})\} \\
g(\{C_{g_5^1}(\overline{E}), C_{h_1^2}(\overline{E})\}) & \rightarrow \{C_{h_2^2}(\overline{E}), C_{h_4^2}(\overline{E})\} \\
f(\{C_{f_3^1}(\overline{E})\}, \{C_{f_3^2}(\overline{E}), C_{h_1^1}(\overline{E})\}) & \rightarrow \{C_{h_2^2}(\overline{E}), C_{h_4^2}(\overline{E})\} \\
& a \rightarrow \{C_{f_3^1}(\overline{E})\}
\end{aligned}$$

## 6 Conclusion

We presented a new and more efficient algorithm for the computation of the equation tree automaton from a regular tree expression by extending the notion of  $k$ -c-continuation from words to trees. We proved that a regular tree expression  $\overline{E}$  can be converted into an equation tree automaton with an  $O(|Q_{\overline{E}}/\sim_e| \cdot |\overline{E}|)$  time and space complexity where  $Q$  is the set of  $k$ -C-Continuations of  $\overline{E}$ .

## References

1. Antimirov, V.: Partial derivatives of regular expressions and finite automaton constructions. *Theoretical computer Science* **155** (1996) 291–319
2. Bruggemann-Klein, A.: Regular expressions into finite automata. *Theoretical computer Science* **120** (1993) 197–213
3. Champarnaud, J.M., Ziadi, D.: From c-continuations to new quadratic algorithms for automaton synthesis. *Intern. J. of Algebra and Computation* **11(6)** (2001) 707–735

4. Champarnaud, J.M., Ziadi, D.: Canonical derivatives, partial derivatives and finite automaton constructions. *Theoretical Computer Science* **289(1)** (2002) 137–163
5. Comon, H., Dauchet, M., Gilleron, R., Jacquemard, F., Lugiez, D., Loding, C., Tison, S., Tommasi, M.: Tree automata techniques and applications. Available on: <http://www.grappa.univ-lille3.fr/tata> (October 2007)
6. Glushkov, V.M.: The abstract theory of automata. *Russian Mathematical Surveys* **16** (1961) 1–53
7. Khorsi, A., Ouardi, F., Ziadi, D.: Fast equation automaton computation. *Journal of Discrete Algorithms* **6** (2008) 433–448
8. Kuske, D., Meinecke, I.: Construction of tree automata from regular expressions. *RAIRO - Theoretical Informatics and Applications* **45** (2011) 347–370
9. Laugerotte, E., Ouali-Sebti, N., Ziadi, D.: From regular tree expression to position tree automaton. *Lecture Notes in Computer Science* **7810** (2013) 395–406
10. Murata, M.: Hedge automata: a formal model for xml schemata. Available on: [http://www.xml.gr.jp/relax/hedge\\_nice.html](http://www.xml.gr.jp/relax/hedge_nice.html) (2000)
11. Nerode, A.: Linear automata transformation. *Proc. Amer. Math. Soc.* **9** (1958) 541–544
12. R. Paige, R.T.: Three partition refinement algorithms. *SIAM Journal on Computing* **16 (6)** (1987) 973–989
13. Revuz, D.: Minimization of acyclic deterministic automata in linear time. *Theoretical Computer Science* **92(1)** (1992) 181–189
14. Trakhtenbrot, B.: Origins and metamorphoses of the trinity: Logic, nets, automata. In *Proceedings, Tenth Annual IEEE Symposium on Logic in Computer Science*. IEEE Computer Society Press (June 1995) 26–29
15. Ziadi, D., Ponty, J.L., Champarnaud, J.M.: Passage d’une expression rationnelle a un automate fini non deterministe. *Bulletin of the Belgian Mathematical Society - Simon Stevin* **4** (1997) 177–203